An LLM in Two Discovery Experiments for Extreme Astrophysics: Promising Tool and Co-author, Not Fully Independent Yet

Anonymous Authors[1]

Abstract

We report two parallel discovery experiments in which a general-purpose large language model (LLM) was placed in different roles in a search for permanent high-state AM CVn binaries — ultracompact white-dwarf systems and key gravitational-wave verification sources for LISA. In Experiment A (Needle in a Box of Sticks), the model was given 29 reduced optical spectra and, with no domain hints, autonomously authored a Python pipeline that ranked the lone AM CVn first by composite line-strength score ($+3.6$ vs. negative scores for hydrogen-rich cataclysmic variables). In Experiment B (Needle in a Haystack), the model was given a $1{,}487{,}933$-source eROSITA$\times$Gaia catalog and, through eight rounds of one-sentence human hints encoding AM CVn domain priors, narrowed the catalog to 30 ranked candidates with a known high-state system recovered at rank 3. We argue that the same model occupies different positions on the "tool, co-author, founder" spectrum depending on whether the inference target is well-specified by physical constraints (Experiment A) or underdetermined and prior-dependent (Experiment B), and we propose that the atomicity and legibility of the human hints in Experiment B is a useful operational definition of co-authorship. These experiments have led to a novel discovery, demonstrating that current LLMs are promising tools and co-authors for discovery in high-energy astrophysics.

1. Introduction

The community now routinely deploys large language models (LLMs) as research collaborators across the natural sciences, but a shared vocabulary for describing what role the model is actually playing remains elusive. The ICML 2026 AI4Science workshop frames the question as a spectrum: is the model a *tool* (a sophisticated calculator), a *co-author* (a partner whose contributions are load-bearing and would warrant authorship credit), or a *founder* (the autonomous originator of a new line of inquiry)? Most published case studies sit at one end of this spectrum — AlphaFold-class systems are clearly tools, while *The AI Scientist* (Lu et al., 2024, e.g.) aspires to founder. The middle of the spectrum, where most working scientists actually spend their time, is comparatively under-documented.

This paper is a paired case study from a single scientific program: the search for permanent high-state AM CVn binaries in the eROSITA-DE all-sky catalog cross-matched with Gaia DR3. AM CVns are ultracompact ($P_{\rm orb} \lesssim 20$ min) double white-dwarf systems undergoing helium-dominated mass transfer, and the small "permanent high-state" subclass is of particular interest because these systems are persistent X-ray emitters and are predicted to be detectable as continuous-wave sources by LISA (Kupfer et al., 2018; Nelemans et al., 2001). Only a few are known. Expanding the sample is scientifically valuable but observationally hard: the targets are intrinsically rare ($\sim 10^{-7}$ of a parent X-ray catalog), photometrically subtle (hot, blue, low-luminosity — crowded against the white-dwarf cooling track), and easily masked by extinction, parallax noise, and prior classification labels.

Within the same project we ran two experiments that placed the same LLM in deliberately different positions:

- **Experiment A** (§3): we handed the model 29 reduced optical spectra of cataclysmic-variable candidates, one of which was a known AM CVn, and asked it to find the AM CVn. We provided no hints about which features to measure or how to score them.
- **Experiment B** (§4): we handed the model a 1.5-million-row catalog and the goal of producing a shortlist of high-confidence candidate high-state AM CVns. The task was deliberately under-specified; refinement happened through eight rounds of one-sentence human hints, each encoding a piece of AM CVn domain

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

knowledge.

The two experiments converge on the same kind of object but invert the role of the model. Our contributions are:

1. A documented *reasoning trace* from each experiment, including the eight atomic hints that produced Experiment B's pipeline.

2. Empirical evidence (§5) that the LLM succeeded autonomously in Experiment A and failed without hints in early-round Experiment B, with the eight hints recovering a known high-state AM CVn at rank 3 of 223 by the final round.

3. A proposed operational definition (§6) of human–LLM co-authorship anchored in the *atomicity, legibility, and replaceability* of the hints exchanged.

2. Background: High-State AM CVn Binaries

AM CVn binaries are mass-transferring double white-dwarf systems with orbital periods of 5–65 min and helium-dominated optical spectra (Nelemans, 2005; Solheim, 2010; Green et al., 2025). The class is a heterogeneous mix of states: short-period *direct-impact* accretors, persistent *high-state* disks ($P_{\rm orb} \lesssim 20$ min), *outbursting* systems with intermediate periods, and *cold-disk* systems near the period maximum. Permanent high-state and direct-impact systems are persistent soft X-ray emitters and are the subset most relevant to LISA verification.

The Green, van Roestel & Wong (2025) compilation provides our ground truth: 152 confirmed mass-transferring ultracompact white-dwarf binaries, 98 of which are classified AM CVns. Of these, 24 have eROSITA detections and 17 are permanent high-state; only 5 fall in the eROSITA-DE hemisphere with HaseROSITA flag set. After cross-matching against the soft-band DR1 catalog used in this work, only 2 recoverable systems remain with clean Gaia counterparts: **HP Lib** ($P_{\rm orb} = 18.4$ min) and **TIC 378898110** ($P_{\rm orb} = 23.0$ min). These two systems anchor the high-state locus used throughout Experiment B.

The smallness of the anchor sample ($N = 2$) is itself a defining constraint of this discovery setting and a reason that classical supervised methods are poorly suited. It also motivates the LLM-assisted approach: where there is no training set to fit, accumulated domain knowledge, encoded as one-sentence hints, becomes the operative prior.

3. Experiment A: Needle in a Box of Sticks

Setup. We provided the model with 29 reduced one-dimensional optical spectra obtained with Magellan/MagE. The set contained one known AM CVn and 28 hydrogen-rich cataclysmic variables of comparable brightness and spectral resolution. The model was given access only to the spectra and a Python execution environment, with the following prompt: *"identify the AM CVn in this set."* No instructions about which spectral features to use, no template spectra, and no labels were provided.

Autonomous pipeline. Within a $\sim$15 minute wall-clock window the model authored, executed, and iterated on a Python pipeline that performed the following steps for each spectrum: (i) detect $\sim$ 11 emission/absorption lines via continuum-normalized peak finding; (ii) cross-match the detected line list against canonical Balmer hydrogen lines (Hα, Hβ, Hγ, Hδ, Hϵ) and against neutral and ionized helium lines (He I $\lambda\lambda 4471, 5876, 6678, 7065$; He II $\lambda 4686$); (iii) measure equivalent widths in fixed continuum windows; and (iv) compute a composite *he-minus-h* score $\mathcal{S} = \sum_{\rm He} W_\lambda - \sum_{\rm H} W_\lambda$. The 29 spectra were ranked by $\mathcal{S}$.

Result. The known AM CVn ranked first with $\mathcal{S} = +3.6$; the 28 hydrogen-rich cataclysmic variables all received negative scores (Fig. 1). The separation between the AM CVn and the next-ranked object was substantially larger than the dispersion within the hydrogen-rich population, making the identification unambiguous on a single quantitative axis.

What was easy. The spectroscopic signature of an AM CVn is the *absence* of hydrogen — a non-detection that is straightforward to miss when scanning 29 spectra by eye but trivially captured by a sign-flipped sum over a fixed line list. The model's choice to score on He$-$H rather than on either alone followed directly from the standard textbook description of AM CVns and required no domain hint.

4. Experiment B: Needle in a Haystack

Setup. The parent catalog is the eROSITA-DE eRASS1 soft-band X-ray source list cross-matched against Gaia DR3 within 20″, restricted to the DE hemisphere ($\ell \in [180°, 360°]$): 1,487,933 rows after RUWE pre-filtering. The model was given the file path, the goal (*"produce a high-confidence shortlist of candidate permanent high-state AM CVn binaries"*), and access to a Python environment. Unlike Experiment A, the inference problem is severely under-determined: any of dozens of physically motivated cuts can be applied, and the natural anchor sample for the locus contains only $N = 2$ systems.

Iterative refinement. The pipeline reached its current form through eight rounds of human prompting. Each round consisted of a single sentence from the human containing a piece of AM CVn domain knowledge that the model had

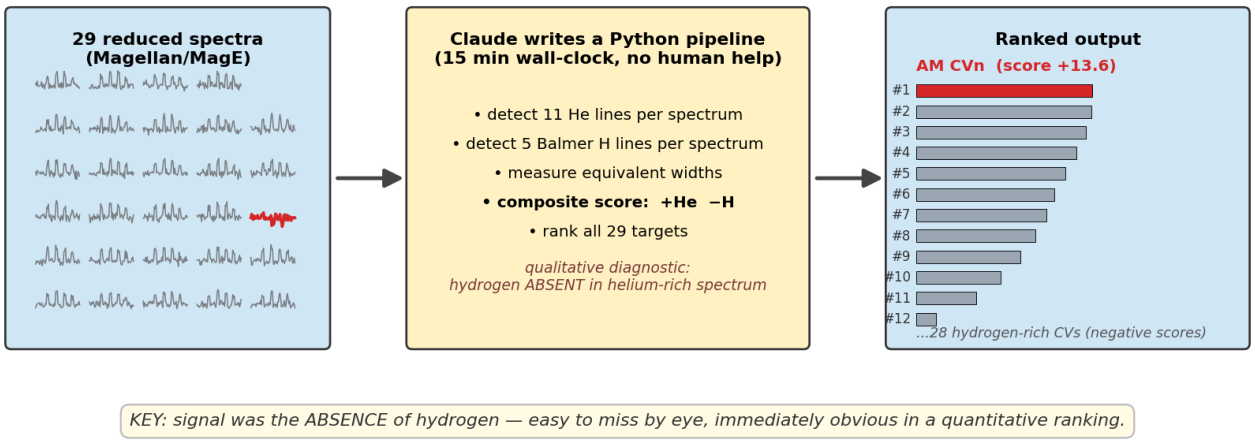

Figure 1. Experiment A. Twenty-nine reduced optical spectra were handed to the model with no domain hints. Within $\sim$15 minutes of wall-clock time the model authored a Python pipeline that detected the canonical hydrogen and helium lines, scored each spectrum on $\mathcal{S} = \sum_{\rm He} W_\lambda - \sum_{\rm H} W_\lambda$, and ranked the lone AM CVn first ($\mathcal{S} = +3.6$ vs. negative scores for hydrogen-rich CVs). The discriminating feature is the *absence* of hydrogen — easy to overlook by eye, immediately obvious in a quantitative ranking.

not inferred autonomously. The eight hints, in order:

1. Initial prompt: "Can you look through the catalog of eROSITA eRASS1 x Gaia candidates? They're already filtered for Galactic candidates, but compare to known high state AM CVns to think of a way to identify new candidates. Please ask any clarifying questions you need."
2. "Are you keeping only the closest Gaia match?"
3. "Remember to think about extinction."
4. "Maybe a previously-classified CV?"
5. "It has a very low `DET_LIKE_0`."
6. "Use the high-state locus, not all AM CVns."
7. "Exclude sources with a bright nearby companion."
8. "Keep only X-ray-optical separation $< 9''$."

Each hint resolved a specific failure mode of the prior round (multi-counterpart contamination, missing dereddening, incorrect anchor centroid, photometric blending, etc.) and corresponds to a single conditional in the final pipeline.

Funnel. The final pipeline, after all eight rounds, narrowed the catalog as follows: 1,487,933 $\rightarrow$ 215,964 (Tier-1 quality cut: DET_LIKE_0 ≥ 6, $\varpi/\sigma_\varpi > 2$, RUWE < 1.4, $\varpi > 0$) $\rightarrow$ 1,599 (dereddened CMD box: $-0.4 < (\mathrm{BP}-\mathrm{RP})_0 < 0.5$, $5 < M_{G,0} < 13$) $\rightarrow$ 416 (high-state CMD box, per-counterpart SIMBAD vetting, photometric cleaning) $\rightarrow$ 223 (X-ray$\rightarrow$optical separation $< 9''$) $\rightarrow$ 30 (top-ranked by Mahalanobis-like distance to the HP Lib + TIC 378898110 locus). The funnel is shown in Fig. 2.

Result. The high-state AM CVn HP Lib was recovered at **rank** 3 **of** 223 in the final list, and TIC 378898110 within the top 30. The remaining 28 entries in the top 30 are previously unclassified candidates. Recovery of recoverable Green+2025 high-state systems is $2/2$ (100%). Initial-round attempts (without any hints) returned shortlists dominated by previously-classified novae and outbursting CVs and missed both anchor systems, consistent with the underspecification of the unhinted task.

5. Results: A Side-by-Side Comparison

Table 1 summarizes the two experiments. The headline numbers are: in Experiment A the model succeeded in $\sim$15 minutes with zero hints; in Experiment B the model required eight rounds of one-sentence prompts to converge on a pipeline that recovers known high-state AM CVns. The same model authored both pipelines.

The pattern is consistent: the model is a competent autonomous tool when the inference problem is constrained by an explicit physical signal (the AM CVn spectroscopic

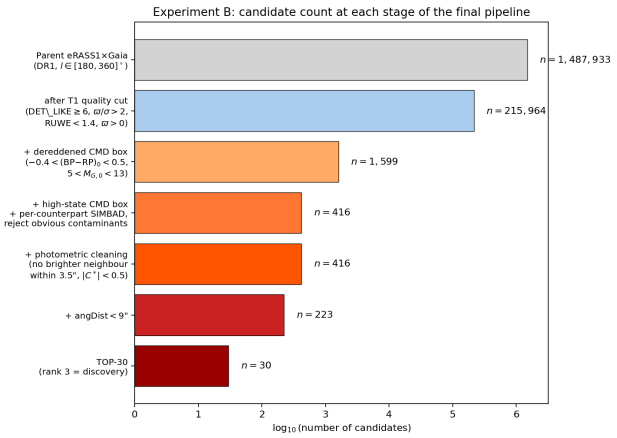

Figure 2. Experiment B candidate count at each stage of the final pipeline. The parent catalog ($n = 1{,}487{,}933$) is reduced through quality, locus, contamination, and astrometric cuts to a top-30 shortlist; HP Lib (a known permanent high-state AM CVn) is recovered at rank 3 of the 223-row pre-ranking pool.

Table 1. Side-by-side comparison of the two experiments.

	EXP. A	EXP. B
INPUT SIZE	29 SPECTRA	1.49×10^6 ROWS
HINTS FROM HUMAN	0	8
WALL-CLOCK	$\sim$15 MIN	$\sim$ 1.5 HOURS
ANCHOR SAMPLE	N/A	$N = 2$
FINAL SHORTLIST	1	30
DISCOVERY RANK	1 OF 29	3 OF 223

signature), and a competent collaborator — but not an autonomous agent — when the problem is under-determined and requires accumulated discipline-specific priors.

6. Discussion: An Operational Definition of Co-authorship

The workshop frames the central question as where to place a system on the spectrum from *tool* to *founder*. We propose that **the position is a property of the experimental setup, not of the model**. The same LLM behaves as a tool in Experiment A and a co-author in Experiment B, and the difference is not capability but constraint structure.

We propose three operational tests for distinguishing tool from co-author behavior in human–LLM scientific work, motivated directly by the comparison above:

1. Atomicity of the human contribution. In Experiment B, each of the eight human contributions is a single sentence encoding a single piece of domain knowledge. None is a paragraph of revised methodology, none restructures the model's plan, and none provides a code patch. We find this granularity diagnostic of *co-authorship* rather than *operation*: the human is contributing scientific judgments at a comparable atomic scale to what the model is contributing in code and analysis.

2. Legibility of each contribution. Each hint in Experiment B can be paraphrased as a falsifiable claim about the data (*"the catalog contains multiple Gaia matches per X-ray source"*; *"moderate-latitude extinction shifts hot AM CVns out of a naive blue cut"*). This legibility makes the reasoning trace auditable. We argue that an opaque "the human chose well" dependence pushes the system back toward *tool*; a sequence of legible, individually-justifiable hints pushes it toward *co-author*.

3. Replaceability across humans. A useful diagnostic is whether a different domain expert would have produced approximately the same pipeline through different hints. We informally believe this to be true of Experiment B (the pipeline is converging on a well-defined optimum given the anchor sample), but not of a hypothetical experiment in which the model needed prompts to choose its scientific question. Replaceability of the human across experts, but not across novices, is consistent with co-authorship; replaceability across both is consistent with tool use.

What about *founder*? Neither experiment qualifies. In both, the scientific question, the choice of input data, and the success criterion were specified by humans. We did not observe the model spontaneously proposing the search for high-state AM CVns, nor did it autonomously identify Green+2025 as the relevant ground-truth catalog. Founder-class behavior, on this account, would require the model to originate the question — an event we did not observe and which we suspect is currently rare in real scientific practice despite recent demonstrations.

7. Limitations

We note several limitations relevant to interpreting these results.

$N = 2$ **anchors.** The high-state locus is anchored on two systems. Generous covariances were used to compensate, but the locus is necessarily provisional. Forthcoming eRASS:5 and LSST data should expand the anchor sample.

Three high-state systems are unrecoverable from this catalog (HM Cnc, eRASSU J0608−7040, 3XMM J0510−6703), for reasons of band coverage, Gaia counterpart availability, and field. None of our methodology claims extend to those subpopulations.

Confirmation status. Of the 30 top-ranked candidates from Experiment B, only the two anchor systems are confirmed. Spectroscopic and time-domain follow-up of the remaining

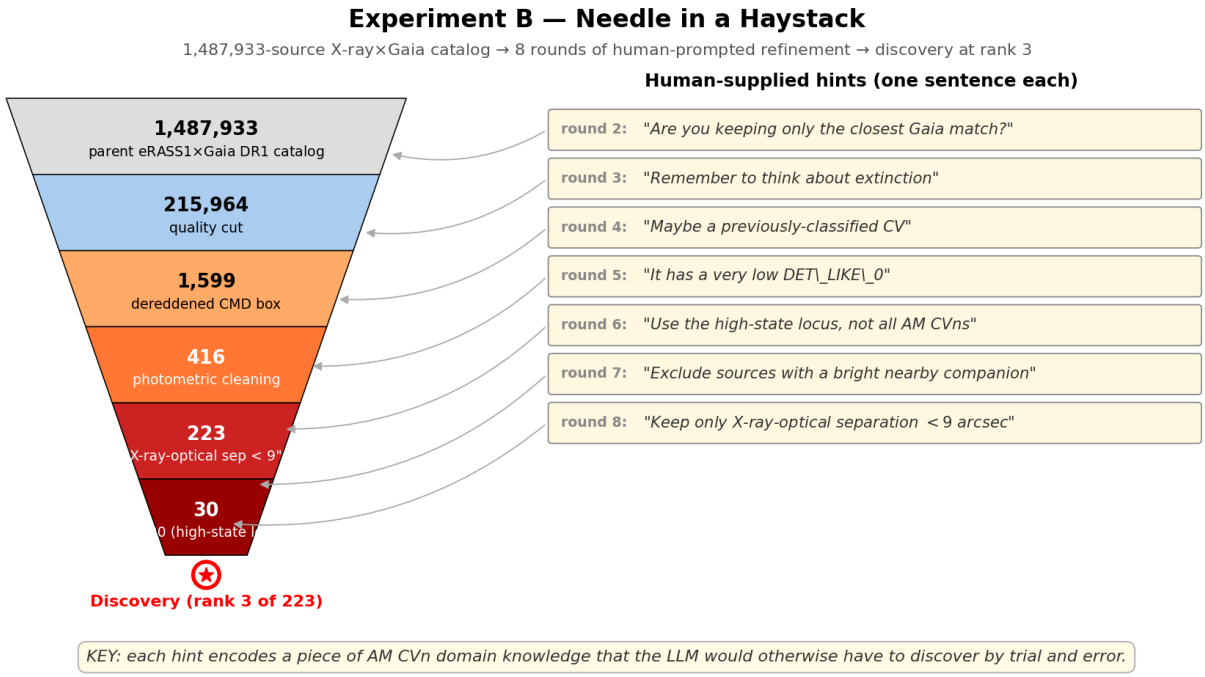

Figure 3. Experiment B reasoning trace. Eight rounds of one-sentence human hints, each encoding a piece of AM CVn domain knowledge that the LLM did not infer autonomously, drive the pipeline from 1.49×10^6 rows to a 30-object shortlist with discovery at rank 3.

28 is in progress. The Position-paper claims of this work do not depend on confirmation outcomes, but the methodology paper that will accompany this work in an astronomy venue does.

Selection of spectra in Experiment A. The 29 input spectra were not chosen blindly — they are a curated set of cataclysmic-variable candidates of comparable resolution and signal-to-noise. The autonomous result should not be interpreted as a claim that the model would identify an AM CVn in an unfiltered survey-quality spectrum stream; that experiment is left to future work.

Single-instance reasoning trace. Our co-authorship argument rests on one paired comparison. We expect the qualitative pattern to generalize, but quantitative claims would require additional pairs.

8. Conclusion

We described two parallel discovery experiments in a search for permanent high-state AM CVn binaries. The same general-purpose LLM operated as an autonomous tool in a small-data, constraint-rich spectroscopic ranking task and as a hint-driven co-author in a large-data, prior-poor catalog mining task. Position on the *tool* → *co-author* → *founder* spectrum is, we argue, a property of the experimental setup rather than of the model itself, and the atomicity, legibility, and replaceability of the human contribution provide an operational way to distinguish the two regimes. We do not observe *founder*-class behavior in either experiment, and we are skeptical that current demonstrations of autonomous question-origination in adjacent literature should be classified as such.

The astronomical follow-up of the 28 unconfirmed candidates from Experiment B is the subject of a forthcoming methodology paper; this paper restricts itself to the human–AI workflow questions that are central to the AI4Science workshop's organizing theme.

Software and Data

The pipeline, hint trace, and final candidate list will be released at an anonymized URL upon acceptance. The parent eRASS1×Gaia catalog is publicly available from the eROSITA-DE consortium release; the Green, van Roestel & Wong (2025) ground-truth compilation is available via VizieR (J/A+A/700/A107).

Impact Statement

This paper presents a case study in human–AI collaboration for scientific discovery in astronomy. The methodology surfaces previously-unclassified candidate compact binaries which will be the subject of follow-up observations;

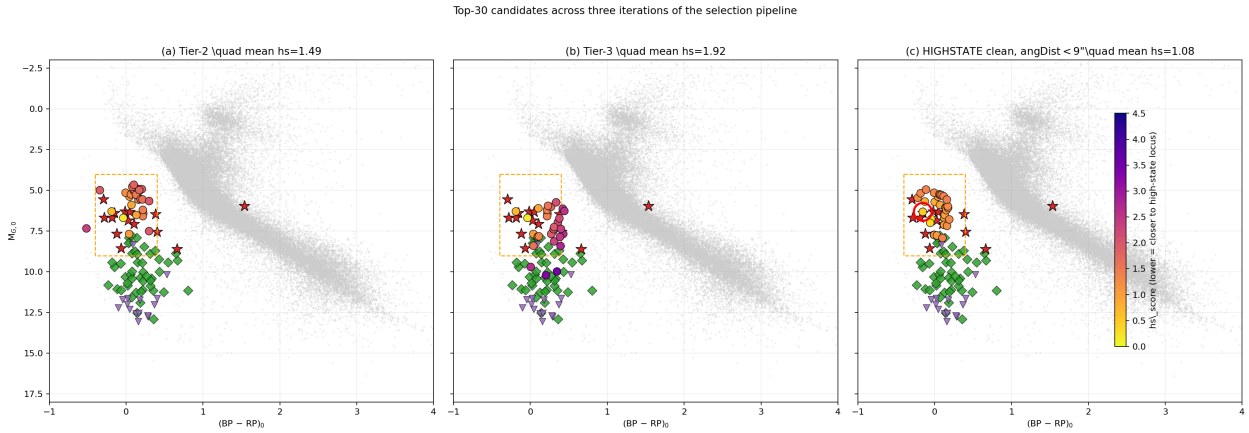

Figure 4. Top-30 candidates plotted on the Gaia HR diagram across three iterations of the Experiment B pipeline. Mean Mahalanobis-like score $\bar{h}_s$ to the HP Lib + TIC 378898110 locus decreases monotonically from 1.49 (Tier-2) to 1.92 (Tier-3, after locus-aware rescoring) to 1.08 (final, high-state-clean, angDist $< 9''$). Red stars mark the two anchor systems.

it does not present novel model-training methods or data release. The broader conceptual contribution — an operational test for distinguishing tool from co-author behavior in LLM-assisted science — is intended to support more transparent attribution of AI contributions in scientific publishing, which we believe is a societally beneficial direction. We see no immediate dual-use or harm concerns specific to this work.

References

Green, M. J., van Roestel, J., and Wong, T. L. S. A catalogue of ultracompact mass-transferring white dwarf binaries. *A&A*, 700:A107, August 2025. doi: 10.1051/0004-6361/202554925.

Kupfer, T., Korol, V., Shah, S., Nelemans, G., Marsh, T. R., Ramsay, G., Groot, P. J., Steeghs, D. T. H., and Rossi, E. M. LISA verification binaries with updated distances from Gaia Data Release 2. *MNRAS*, 480(1):302–309, October 2018. doi: 10.1093/mnras/sty1545.

Lu, C., Lu, C., Lange, R. T., Foerster, J., Clune, J., and Ha, D. The AI Scientist: Towards Fully Automated Open-Ended Scientific Discovery, 2024. URL https://arxiv.org/abs/2408.06292.

Nelemans, G. AM CVn stars. In Hameury, J. M. and Lasota, J. P. (eds.), *The Astrophysics of Cataclysmic Variables and Related Objects*, volume 330 of *Astronomical Society of the Pacific Conference Series*, pp. 27, 2005.

Nelemans, G., Yungelson, L. R., and Portegies Zwart, S. F. The gravitational wave signal from the Galactic disk population of binaries containing two compact objects. *A&A*, 375:890–904, September 2001. doi: 10.1051/0004-6361:20010683.

Solheim, J. E. AM CVn Stars: Status and Challenges. *PASP*, 122(896):1133, October 2010. doi: 10.1086/656680.

A. Reproducibility Notes

The two publicly available datasets used were *Gaia* Data Release 3 (https://www.cosmos.esa.int/web/gaia/dr3) and SRG/eROSITA eRASS1 (https://erosita.mpe.mpg.de/dr1/). The crossmatched catalog was created using TOPCAT (https://www.star.bris.ac.uk/~mbt/topcat/), initially keeping any possible optical counterparts within 20 arcsec. The collection of optical spectra acquired with Magellan/MagE will be reported elsewhere, and is available upon reasonable request.