VERIFIABLE RESEARCH AGENTS FOR CHRONIC KIDNEY DISEASE: A RULE-ENGINE HARNESS WITH EXPERTS IN THE LOOP

RESEARCH PLAN • DRAFT FOR DISCUSSION

Hayat Lab, Uniklinik RWTH Aachen
Draft for internal discussion; author list to be confirmed

3 November 2026

ABSTRACT

Research agents built on large language models already analyse chronic kidney disease (CKD) data and propose drug targets, yet their conclusions arrive as free text that nobody checks systematically, and expert time is scarce. We propose a harness that wraps any such agent and makes its conclusions verifiable. Each conclusion is turned into a structured claim, generated from the agent's code output, and checked by a rule engine on three layers: domain knowledge, data validity and analysis code. A broken hard rule blocks a claim; a broken soft rule flags it for review. The harness selects the most informative flagged claims for experts and turns their verdicts into versioned rules, so that it improves from round to round. We will evaluate it with planted errors, ablations, learning curves and the replication of checked claims in independent CKD cohorts. This document describes the design; no results are reported yet.

Keywords LLM agents · verification · rule engine · human in the loop · chronic kidney disease · single-cell omics

1 Introduction

Motivation. Research agents now work much like a graduate student: they read the literature, write and run analysis code, and report conclusions. They have begun to be applied to CKD. NephroAIX ranks kidney targets over a knowledge graph built from single-cell data [1], and Robin proposed ten CKD drug candidates from the literature alone [2]. Neither result has been checked against data. In drug discovery, a wrong target wastes all the work that follows.

Problem. Agents also make substantive errors. They fabricate data when it is missing [3], reach conclusions that do not survive small perturbations [4], and treat thousands of cells from one patient as independent samples [5]. Kidney data add traps of their own, such as cohorts that appear independent but share patients (Appendix A). Experts can catch such errors, but they cannot read every report.

Gap. Many agents can analyse such data; what is missing is the layer around them. Current evaluations grade reports with language models or rubrics [6], and related verification work checks reasoning, statistics or knowledge separately and outside omics [7–9]. To our knowledge, no harness checks an agent's claims across domain knowledge, data and code on one engine, involves experts only where the rules call for it, and learns new rules from their verdicts.

Contributions. We contribute (i) a harness that turns an agent's conclusions into structured claims and checks them on three layers with one declarative rule engine; (ii) active selection of claims for expert review, with expert verdicts converted into versioned rules; (iii) a CKD instantiation with a released rule base; and (iv) an evaluation design based on planted errors, ablation, learning curves and replication in independent CKD cohorts.

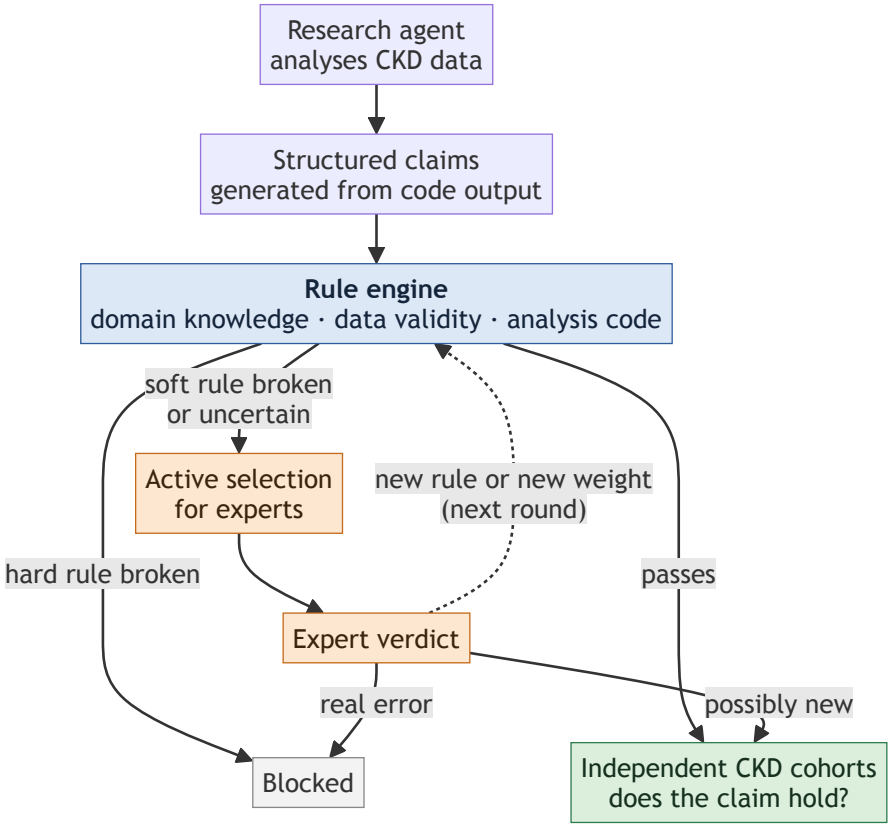


Figure 1: The harness. Claims generated from the agent's code output pass through the rule engine (blue). Claims that break a soft rule or are uncertain are selected for experts (orange), whose verdicts update the rules for the next round (dotted arrow). Claims that pass, and claims that experts judge possibly new, are tested on independent CKD cohorts (green).

2 The harness

Figure 1 shows the loop. Any agent analyses the CKD data; the harness checks each claim on the way and calls on experts only where needed.

Structured claims. Every conclusion becomes a record with fixed fields: cell type, gene, direction and size of the effect, the groups compared, the unit of analysis, the number of people per group, the type of evidence (association or intervention) and the code output behind each number. The record is generated from the code output rather than written by the language model, which could otherwise phrase a claim so that it passes.

Three layers. The engine checks each claim on three layers, defined by where the facts come from (Table 1). Each layer is an established field: knowledge-based checking, data validation [10] and program analysis [11]. Data-driven discovery is known to require all three kinds of expertise [12], but they have not been combined in one engine over agent claims. The engine reads the code and the data, not only the report; otherwise an agent could declare a per-patient analysis while computing per cell [13].

Hard and soft rules. A broken hard rule, such as "statistics per patient", blocks a claim. A broken soft rule, such as "marker genes should match the cell type", only flags it, because a genuine new finding may contradict current knowledge. The rules run on the answer-set-programming system clingo; clingo-explaid returns the smallest set of conflicting rules, which is what the expert sees [14, 15]. The full rule list and tools are in Appendix B.

Experts in the loop. Experts never read whole reports. The harness ranks flagged claims by the size of their rule conflicts and the uncertainty of the soft rules involved, and shows each selected claim together with the rules it breaks

Table 1: The three layers of the rule engine.

Layer	What it checks	CKD examples	Facts come from
Domain knowledge	does the claim agree with CKD biology	marker genes match the cell type; direction agrees with known injury states; association not stated as cause	Cell Ontology, HuBMAP kidney table, KPMP atlas
Data validity	are the data reliable and used correctly	no test-cohort patient in the discovery cohort; doublets, ambient RNA, low-quality cells; enough people per group	the data and their metadata
Analysis code	are statistics and computations right	statistics per patient, not per cell; no testing of clusters on the data that defined them; multiple-testing correction; numbers traced to output	the agent's code and data flow

and the code that produced it. This selection is compared against random selection, which is needed because selecting by an agent's own confidence can do worse than random [16]. A verdict of "real error" can become a new rule; "rule does not apply" lowers the weight of a soft rule. Every rule records its source and carries a retirement condition, since uncurated rules can perform worse than none [17]. The rule base is frozen within an evaluation round and updated only between rounds (Appendix C).

3 Evaluation

Four evaluations run from the most controlled to the closest to practice. *Planted errors:* someone other than the rule authors plants known errors (per-cell statistics, mixed cohorts, mislabelled cell types, invented numbers) into real CKD analyses, and we count per layer what is caught and what is falsely flagged. *Ablation:* the same agents run with and without the harness, and with one layer removed at a time, each with a closed and an open-weight model and at least three reruns, so that the effect of the harness can be compared with that of a new model generation. *Learning curves:* after each round we measure errors caught per expert-minute and expert time per claim. *Replication, the endpoint:* claims that passed and claims that were flagged are tested on CKD cohorts the agent has never seen; if passing claims hold more often, the harness can stop unreliable conclusions before a second cohort exists. Appendix E describes how the evaluation data are built.

4 CKD data and first phase

The agent works on the KPMP kidney atlas [18]; independent cohorts from other groups [19–21] provide the test data, and patients shared between cohorts are removed before any split (Appendix F). Answers are recomputed from the released data, never taken from the papers' written conclusions.

The first phase runs for six months, from November 2026 to April 2027, with the postdoc and one master's student (Appendix H). The student's thesis is the rule engine and the selection of claims for experts; the postdoc prepares the CKD evaluation data, organises the expert rounds and analyses the results. The pilot uses one discovery and two test cohorts, two agents, two base models, about 20 questions and three reruns.

References

[1] Sikander Hayat, Markus Joppich, and Rafael Kramann. NephroAIX: Large Language Model (LLM)-Guided Knowledge Graph for Explainable Target Discovery in Human Kidney Diseases Using Single-Cell Data: FR-OR014. *Journal of the American Society of Nephrology*, 36(10S), 2025. ISSN 1533-3450. doi: 10.1681/asn.20254q7nxd06. URL http://dx.doi.org/10.1681/ASN.20254q7nxd06.

[2] Ali Essam Ghareeb, Benjamin Chang, Ludovico Mitchener, Angela Yiu, Caralyn J. Szostkiewicz, Jon M. Laurent, Muhammed T. Razzak, Andrew D. White, Michaela M. Hinks, and Samuel G. Rodriques. Robin: A multi-agent system for automating scientific discovery, 2025. URL https://arxiv.org/abs/2505.13400.

[3] Zonglin Yang, Xingtong Liu, and Xinyan Xu. SciIntegrity-Bench: A Benchmark for Evaluating Academic Integrity in AI Scientist Systems, 2026. URL https://arxiv.org/abs/2605.10246.

[4] Zachary T. Rewolinski, Austin V. Zane, Hao Huang, Chandan Singh, Chenglong Wang, Jianfeng Gao, and Bin Yu. Sanity Checks for Agentic Data Science, 2026. URL https://arxiv.org/abs/2604.11003.

[5] Samuel Alber, Bowen Chen, Eric Sun, Alina Isakova, Aaron J. Wilk, and James Zou. CellVoyager: AI CompBio agent generates new insights by autonomously analyzing biological data. *Nature Methods*, 23(4):749–759, March 2026. ISSN 1548-7105. doi: 10.1038/s41592-026-03029-6. URL http://dx.doi.org/10.1038/s41592-026-03029-6.

[6] Yuanhao Qu et al. BiomniBench: Process-level Evaluation of LLM Agents for Real-world Biomedical Research. May 2026. doi: 10.64898/2026.05.12.724604. URL http://dx.doi.org/10.64898/2026.05.12.724604.

[7] Bobo Li, Hao Fei, Tianjie Ju, Mong-Li Lee, and Wynne Hsu. OmniScientist: An Omni-Modal Omni-Discipline AI Scientist, 2026. URL https://arxiv.org/abs/2608.13558.

[8] Yu Feng, Nathaniel Weir, Kaj Bostrom, Sam Bayless, Darion Cassel, Sapana Chaudhary, Benjamin Kiesl-Reiter, and Huzefa Rangwala. VeriCoT: Neuro-symbolic Chain-of-Thought Validation via Logical Consistency Checks, 2025. URL https://arxiv.org/abs/2511.04662.

[9] John Seon Keun Yi, Joshua R. Minot, and Dokyun Lee. GRACE: Graph-Grounded Reflective Agent Copilot Engine for Expert-in-the-Loop Knowledge Expansion, 2026. URL https://arxiv.org/abs/2609.04442.

[10] Sebastian Schelter, Dustin Lange, Philipp Schmidt, Meltem Celikel, Felix Biessmann, and Andreas Grafberger. Automating large-scale data quality verification. *Proceedings of the VLDB Endowment*, 11(12):1781–1794, August 2018. ISSN 2150-8097. doi: 10.14778/3229863.3229867. URL http://dx.doi.org/10.14778/3229863.3229867.

[11] Chenyang Yang, Rachel A Brower-Sinning, Grace Lewis, and Christian KÄStner. Data Leakage in Notebooks: Static Detection and Better Processes. In *Proceedings of the 37th IEEE/ACM International Conference on Automated Software Engineering*, ASE '22, page 1–12. ACM, October 2022. doi: 10.1145/3551349.3556918. URL http://dx.doi.org/10.1145/3551349.3556918.

[12] Ken Gu et al. BLADE: Benchmarking Language Model Agents for Data-Driven Science, 2025. URL https://arxiv.org/abs/2408.09667.

[13] Jordan W. Squair et al. Confronting false discoveries in single-cell differential expression. *Nature Communications*, 12(1), 2021. ISSN 2041-1723. doi: 10.1038/s41467-021-25960-2. URL http://dx.doi.org/10.1038/s41467-021-25960-2.

[14] Potassco. clingo: a grounder and solver for answer set programming. https://github.com/potassco/clingo, 2026. MIT licence.

[15] Potassco. clingo-explaid: explanations for answer set programs. https://github.com/potassco/clingo-explaid, 2026. MIT licence.

[16] Cesare Zavattari, Alessandro Tommasi, and Giuseppe Prencipe. One Human, N Agents: Audit-Budget Allocation for LLM Agent Fleets under Miscalibrated, Correlated Confidence, 2026. URL https://arxiv.org/abs/2607.28317.

[17] Yanwei Cui, Xing Zhang, Yulong Zhang, Li Shao, Xiaofeng Shi, Guanghui Wang, and Peiyang He. Closing the Feedback Loop: From Experience Extraction to Insight Governance in Verbal Reinforcement Learning, 2026. URL https://arxiv.org/abs/2606.17591.

[18] Blue B. Lake et al. An atlas of healthy and injured cell states and niches in the human kidney. *Nature*, 619(7970):585–594, July 2023. ISSN 1476-4687. doi: 10.1038/s41586-023-05769-3. URL http://dx.doi.org/10.1038/s41586-023-05769-3.

[19] Amin Abedini et al. Single-cell multi-omic and spatial profiling of human kidneys implicates the fibrotic microenvironment in kidney disease progression. *Nature Genetics*, 56(8):1712–1724, 2024. ISSN 1546-1718. doi: 10.1038/s41588-024-01802-x. URL http://dx.doi.org/10.1038/s41588-024-01802-x.

[20] Parker C. Wilson, Yoshiharu Muto, Haojia Wu, Anil Karihaloo, Sushrut S. Waikar, and Benjamin D. Humphreys. Multimodal single cell sequencing implicates chromatin accessibility and genetic background in diabetic kidney disease progression. *Nature Communications*, 13(1), 2022. ISSN 2041-1723. doi: 10.1038/s41467-022-32972-z. URL http://dx.doi.org/10.1038/s41467-022-32972-z.

[21] Christoph Kuppe et al. Decoding myofibroblast origins in human kidney fibrosis. *Nature*, 589(7841): 281–286, 2020. ISSN 1476-4687. doi: 10.1038/s41586-020-2941-1. URL http://dx.doi.org/10.1038/s41586-020-2941-1.

[22] FutureHouse. Robin example notebook: chronic kidney disease (robin_CKD.ipynb). https://github.com/Future-House/robin/blob/main/examples/robin_CKD.ipynb, 2025. GitHub repository Future-House/robin, accessed 2026-11-02.

[23] Nathan Nakatsuka, Drew Adler, Longda Jiang, Austin Hartman, Evan Cheng, Eric Klann, and Rahul Satija. Improving reproducibility of differentially expressed genes in single-cell transcriptomic studies of neurodegenerative diseases through meta-analysis. *Nature Communications*, 16(1), 2025. ISSN 2041-1723. doi: 10.1038/s41467-025-62579-z. URL http://dx.doi.org/10.1038/s41467-025-62579-z.

[24] Sirui Ding, Sanchita Bhattacharya, Minnie M. Sarwal, Marina Sirota, and Atul J. Butte. Transplant-Agents: A Multi-Agent Artificial Intelligence Framework for Reproducibility Assessment of Post-Transplant Risk Prediction and Rejection Biomarkers. *bioRxiv*, 2025. doi: 10.1101/2025.07.10.664265. URL http://dx.doi.org/10.1101/2025.07.10.664265.

[25] Konstantinos Stasinos et al. An Integrated Atlas of the Human Kidney Spanning Health and Disease. *bioRxiv*, 2026. doi: 10.64898/2026.08.12.744548. URL http://dx.doi.org/10.64898/2026.08.12.744548.

[26] Yoshiharu Muto, Parker C. Wilson, Nicolas Ledru, Haojia Wu, Henrik Dimke, Sushrut S. Waikar, and Benjamin D. Humphreys. Single cell transcriptional and chromatin accessibility profiling redefine cellular heterogeneity in the adult human kidney. *Nature Communications*, 12(1), 2021. ISSN 2041-1723. doi: 10.1038/s41467-021-22368-w. URL http://dx.doi.org/10.1038/s41467-021-22368-w.

[27] Parker C. Wilson, Haojia Wu, Yuhei Kirita, Kohei Uchimura, Nicolas Ledru, Helmut G. Rennke, Paul A. Welling, Sushrut S. Waikar, and Benjamin D. Humphreys. The single-cell transcriptomic landscape of early human diabetic nephropathy. *Proceedings of the National Academy of Sciences*, 116(39):19619–19625, 2019. ISSN 1091-6490. doi: 10.1073/pnas.1908706116. URL http://dx.doi.org/10.1073/pnas.1908706116.

[28] Bernhard Dumoulin et al. Spatial atlas of diabetic kidney disease reveals a B cell-rich subgroup. *Nature*, 653 (8116):1158–1169, April 2026. ISSN 1476-4687. doi: 10.1038/s41586-026-10363-4. URL http://dx.doi.org/10.1038/s41586-026-10363-4.

[29] Benedikt Boecking, Willie Neiswanger, Eric Xing, and Artur Dubrawski. Interactive Weak Supervision: Learning Useful Heuristics for Data Labeling, 2021. URL https://arxiv.org/abs/2012.06046.

[30] Cheng-Yu Hsieh, Jieyu Zhang, and Alexander Ratner. Nemo: Guiding and Contextualizing Weak Supervision for Interactive Data Programming, 2022. URL https://arxiv.org/abs/2203.01382.

[31] Pouya Ghiasnezhad Omran, Michael Zimmermann, Duncan Cambridge, Ashmita Kapoor, and Tanya Dixit. Agent Gym: A Framework for Continuous Evaluation and Evolution of LLM Agents Through Human-in-the-Loop Feedback, 2026. URL https://arxiv.org/abs/2608.15591.

[32] Karen Sargsyan. Structural Enforcement of Statistical Rigor in AI-Driven Discovery: A Functional Architecture, 2026. URL https://arxiv.org/abs/2511.06701.

[33] Alexi Turcotte and Neev Nirav Mehta. The Fault in our Stats. https://reallytg.github.io/files/papers/ase25_statlint_final_final.pdf, 2025. ASE 2025; introduces the STAT-LINT checker.

[34] Xue Li, Hongqian Xia, Daniel Gomm, Aécio Santos, Cornelius Wolff, and Madelon Hulsebos. The Silent Failures in Agentic Data Science, 2026. URL https://arxiv.org/abs/2610.14115.

[35] Zhizheng Wang, Qiao Jin, Chih-Hsuan Wei, Shubo Tian, Po-Ting Lai, Qingqing Zhu, Chi-Ping Day, Christina Ross, Robert Leaman, and Zhiyong Lu. GeneAgent: self-verification language agent for gene-set analysis using domain databases. *Nature Methods*, 22(8):1677–1685, 2025. ISSN 1548-7105. doi: 10.1038/s41592-025-02748-6. URL http://dx.doi.org/10.1038/s41592-025-02748-6.

[36] Shreya Johri et al. Evaluating agentic AI for biological discovery in autonomous and copilot settings. *bioRxiv*, 2026. doi: 10.64898/2026.06.04.729919. URL http://dx.doi.org/10.64898/2026.06.04.729919.

[37] Kaleigh F. Roberts et al. OpenScientist: evaluating an open agentic AI co-scientist to accelerate biomedical discovery. *medRxiv*, 2026. doi: 10.64898/2026.03.15.26348338. URL http://dx.doi.org/10.64898/2026.03.15.26348338.

[38] Manik Garg, Xu Li, Pablo Moreno, Irene Papatheodorou, Yuelong Shu, Alvis Brazma, and Zhichao Miao. Meta-analysis of COVID-19 single-cell studies confirms eight key immune responses. *Scientific Reports*, 11(1), 2021. ISSN 2045-2322. doi: 10.1038/s41598-021-00121-z. URL http://dx.doi.org/10.1038/s41598-021-00121-z.

[39] Kip D. Zimmerman, Mark A. Espeland, and Carl D. Langefeld. A practical solution to pseudoreplication bias in single-cell studies. *Nature Communications*, 12(1), 2021. ISSN 2041-1723. doi: 10.1038/s41467-021-21038-1. URL http://dx.doi.org/10.1038/s41467-021-21038-1.

[40] Alan E. Murphy and Nathan G. Skene. A balanced measure shows superior performance of pseudobulk methods in single-cell RNA-sequencing analysis. *Nature Communications*, 13(1), 2022. ISSN 2041-1723. doi: 10.1038/s41467-022-35519-4. URL http://dx.doi.org/10.1038/s41467-022-35519-4.

[41] Lucy L. Gao, Jacob Bien, and Daniela Witten. Selective Inference for Hierarchical Clustering. *Journal of the American Statistical Association*, 119(545):332–342, 2022. ISSN 1537-274X. doi: 10.1080/01621459.2022.2116331. URL http://dx.doi.org/10.1080/01621459.2022.2116331.

[42] Anna Neufeld, Lucy L. Gao, Joshua Popp, Alexis Battle, and Daniela Witten. Inference after latent variable estimation for single-cell RNA sequencing data, 2022. URL https://arxiv.org/abs/2207.00554.

[43] HuBMAP. ASCT+B table for the kidney, version 1.8. https://cdn.humanatlas.io/digital-objects/asct-b/kidney/latest/assets/asct-b-vh-kidney.csv, 2026. CC BY 4.0, dated 2026-06-15.

[44] Katy Börner et al. Anatomical structures, cell types and biomarkers of the Human Reference Atlas. *Nature Cell Biology*, 23(11):1117–1128, 2021. ISSN 1476-4679. doi: 10.1038/s41556-021-00788-6. URL http://dx.doi.org/10.1038/s41556-021-00788-6.

[45] Hongbo Liu et al. Kidney multiome-based genetic scorecard reveals convergent coding and regulatory variants. *Science*, 387(6734), 2025. ISSN 1095-9203. doi: 10.1126/science.adp4753. URL http://dx.doi.org/10.1126/science.adp4753.

[46] RDFLib. pySHACL: a Python validator for SHACL. https://github.com/RDFLib/pySHACL, 2026. Apache-2.0 licence.

[47] DTAI, KU Leuven. ProbLog: probabilistic logic programming. https://github.com/ML-KULeuven/problog, 2026. Apache-2.0 licence.

[48] MARTIN GEBSER, TORSTEN SCHAUB, SVEN THIELE, and PHILIPPE VEBER. Detecting inconsistencies in large biological networks with answer set programming. *Theory and Practice of Logic Programming*, 11(2-3):323–360, 2011. ISSN 1475-3081. doi: 10.1017/s1471068410000554. URL http://dx.doi.org/10.1017/S1471068410000554.

[49] Sven Thiele, Luca Cerone, Julio Saez-Rodriguez, Anne Siegel, Carito Guziołowski, and Steffen Klamt. Extended notions of sign consistency to relate experimental data to signaling and regulatory network topologies. *BMC Bioinformatics*, 16(1), 2015. ISSN 1471-2105. doi: 10.1186/s12859-015-0733-7. URL http://dx.doi.org/10.1186/s12859-015-0733-7.

[50] Shreya Shankar, J. D. Zamfirescu-Pereira, Björn Hartmann, Aditya G. Parameswaran, and Ian Arawjo. Who Validates the Validators? Aligning LLM-Assisted Evaluation of LLM Outputs with Human Preferences, 2024. URL https://arxiv.org/abs/2404.12272.

[51] Kexin Huang et al. Autonomous biomedical research with an artificial intelligence agent. *Science*, 393(6813), 2026. ISSN 1095-9203. doi: 10.1126/science.adz4351. URL http://dx.doi.org/10.1126/science.adz4351.

[52] Henry E. Miller, Matthew Greenig, Benjamin Tenmann, and Bo Wang. BioML-bench: Evaluation of AI Agents for End-to-End Biomedical ML. September 2025. doi: 10.1101/2025.09.01.673319. URL http://dx.doi.org/10.1101/2025.09.01.673319.

[53] Mike A. Merrill et al. Terminal-Bench: Benchmarking Agents on Hard, Realistic Tasks in Command Line Interfaces, 2026. URL https://arxiv.org/abs/2601.11868.

[54] Jie Huang, Xinyun Chen, Swaroop Mishra, Huaixiu Steven Zheng, Adams Wei Yu, Xinying Song, and Denny Zhou. Large Language Models Cannot Self-Correct Reasoning Yet, 2024. URL https://arxiv.org/abs/2310.01798.

[55] Zhibin Gou, Zhihong Shao, Yeyun Gong, Yelong Shen, Yujiu Yang, Nan Duan, and Weizhu Chen. CRITIC: Large Language Models Can Self-Correct with Tool-Interactive Critiquing, 2024. URL https://arxiv.org/abs/2305.11738.

[56] Xiangyi Li et al. SkillsBench: Benchmarking How Well Agent Skills Work Across Diverse Tasks, 2026. URL https://arxiv.org/abs/2602.12670.

[57] Mosab Alfaqeeh, Omar Mohammed Horani, Mou'th Hourani, and Hashem Alfaqeeh. Beyond fabrication: Misattributed numbers in papers written by AI research agents, 2026. URL https://arxiv.org/abs/2610.38691.

[58] Yufeng Du, Minyang Tian, Srikanth Ronanki, Subendhu Rongali, Sravan Bodapati, Aram Galstyan, Azton Wells, Roy Schwartz, Eliu A Huerta, and Hao Peng. Context Length Alone Hurts LLM Performance Despite Perfect Retrieval, 2025. URL https://arxiv.org/abs/2510.05381.

[59] Pengcheng Jiang, Zhiyi Shi, Kelly Hong, Xueqiang Xu, Jiashuo Sun, Jimeng Sun, Hammad Bashir, and Jiawei Han. Harness-1: Reinforcement Learning for Search Agents with State-Externalizing Harnesses, 2026. URL https://arxiv.org/abs/2606.02373.

[60] yibie. Jev engineering: for coding agents (Chinese translation). https://github.com/yibie/jev-engineering-zh, 2026. Translation of a third-party compilation of design notes by Diogo Almeida (TypeSafe), September 2026; engineering note, not peer reviewed, no experiments.

[61] Xun Liu, Yi Evie Zhang, Vira Kasprova, Parisa Rabbani, Pardis Sadat Zahraei, Tianyu Zhang, Ali Ebrahimpour-Boroojeny, and Varun Chandrasekaran. AgentAbstain: Do LLM Agents Know When Not to Act?, 2026. URL https://arxiv.org/abs/2607.10059.

[62] Stephan Rabanser, Sayash Kapoor, Peter Kirgis, Kangheng Liu, Saiteja Utpala, and Arvind Narayanan. Towards a Science of AI Agent Reliability, 2026. URL https://arxiv.org/abs/2602.16666.

[63] Yubin Kim et al. Towards a Science of Scaling Agent Systems, 2026. URL https://arxiv.org/abs/2512.08296.

[64] Konstantin A. Klötzer et al. Analysis of individual patient pathway coordination in a cross-species single-cell kidney atlas. *Nature Genetics*, 57(8):1922–1934, August 2025. ISSN 1546-1718. doi: 10.1038/s41588-025-02285-0. URL http://dx.doi.org/10.1038/s41588-025-02285-0.

[65] Petter Bjornstad et al. SGLT2 inhibition modulates metabolic, vascular, and inflammatory molecular markers in the kidney in youth with type 1 diabetes. *Science Translational Medicine*, 18(859), 2026. ISSN 1946-6242. doi: 10.1126/scitranslmed.aee1005. URL http://dx.doi.org/10.1126/scitranslmed.aee1005.

[66] Haojia Wu, Romer Gonzalez Villalobos, Xiang Yao, Dermot Reilly, Tao Chen, Matthew Rankin, Eugene Myshkin, Matthew D. Breyer, and Benjamin D. Humphreys. Mapping the single-cell transcriptomic response of murine diabetic kidney disease to therapies. *Cell Metabolism*, 34(7):1064–1078.e6, 2022. ISSN 1550-4131. doi: 10.1016/j.cmet.2022.05.010. URL http://dx.doi.org/10.1016/j.cmet.2022.05.010.

[67] Amin Abedini et al. Single-cell transcriptomics and chromatin accessibility profiling elucidate the kidney-protective mechanism of mineralocorticoid receptor antagonists. *Journal of Clinical Investigation*, 134(1), 2024. ISSN 1558-8238. doi: 10.1172/jci157165. URL http://dx.doi.org/10.1172/JCI157165.

[68] Michael S. Balzer, Jianfu Zhou, Amin Abedini, Ziyuan Ma, Yanjuan Hou, Tejaswini Yadav, Manuel Grundmann, Mira Pavkovic, and Katalin Susztak. Single-Cell Resolution Drug Effects of Renin-Angiotensin-Aldosterone Blockade in ZSF1 Rat Diabetic Kidney Disease. *Journal of the American Society of Nephrology*, 37(7):1438–1456, 2026. ISSN 1533-3450. doi: 10.1681/asn.0000000995. URL http://dx.doi.org/10.1681/ASN.0000000995.

[69] Katherine R. Tuttle et al. Effects of semaglutide on kidney disease in type 2 diabetes: a randomized placebo-controlled trial. *Nature Medicine*, 2026. ISSN 1546-170X. doi: 10.1038/s41591-026-04674-2. URL http://dx.doi.org/10.1038/s41591-026-04674-2.

[70] Jülich Supercomputing Centre. JUPITER technical overview. https://www.fz-juelich.de/en/jsc/jupiter/tech, 2026. Accessed 2026-11-02.

[71] Daniel P. Jeong, Saurabh Garg, Zachary C. Lipton, and Michael Oberst. Medical Adaptation of Large Language and Vision-Language Models: Are We Making Progress?, 2024. URL https://arxiv.org/abs/2411.04118.

A Background and related work

AI already proposes kidney targets, without validation. NephroAIX ranks kidney-disease targets with a language model over a knowledge graph built from KPMP single-cell data; its abstract reports no validation [1]. In its chronic kidney disease example, Robin searched the literature only and proposed ten drug candidates, most acting on the NLRP3 inflammasome pathway, ranked by pairwise model comparisons [2, 22]. Neither set of results has been tested on data or in experiments.

Agents make substantive errors when they analyse data. A single-cell analysis roughly proceeds from loading data and quality control, through cell-type assignment and comparison of patients with controls, to statistical testing and a report; this is how CellVoyager works, for example [5]. Other groups have shown that agents fail in these steps. In scenarios where only admitting failure is correct, 34.2% of 231 runs across seven frontier models showed an integrity problem, and in missing-data scenarios all seven models synthesised data rather than declare the task infeasible [3]. When OpenAI Codex reanalysed 11 real datasets under reasonable perturbations, the affirmative conclusion of a single run was not well supported in six of them, and the agent's self-reported confidence did not match the stability of its conclusions [4]. In a COVID-19 single-cell analysis, CellVoyager tested cells as independent samples and switched to donor-level pseudobulk only when the original authors asked [5].

Single-cell findings often fail in another set of patients. Differentially expressed genes found in one single-nucleus study separate cases from controls in other studies of the same disease with a mean AUC of only 0.68 for Alzheimer's disease and 0.77 for Parkinson's disease [23]. If findings made by people fare like this, what an agent "finds" in one dataset may well hold only there.

Why this evidence is not from kidney. To our knowledge, neither problem has been measured in kidney. Agent work in kidney is limited to generating targets without validation [1, 2], prediction within a single transplant cohort [24] and cell annotation; and kidney atlases harmonise annotations without testing whether findings replicate [25]. The absence of kidney evidence is itself the gap.

The same traps are visible in kidney data. Checking public metadata (2026-11-02), we found that kidney cohorts that look independent share patients: the controls of Wilson et al. [20] are donors of Muto et al. [26], part of its diabetic kidney disease samples come from Wilson et al. [27], and Dumoulin et al. [28] and Abedini et al. [19] carry identical patient identifiers. The same injured proximal tubule state is described differently: Wilson et al. [20] find VCAM1$^+$ cells more frequent in diabetic kidney disease, with a proportion that "varied widely by donor" and only five to seven patients, whereas in Abedini et al. [19] VCAM1$^+$ injured cells also occur in healthy samples and the CKD-enriched subtype is HAVCR1$^+$. General agents do not know such traps; kidney rules are needed to catch them.

Closest work. No single part of the harness is new. Claims have been checked against an evidence graph with only important and uncertain claims sent to experts, improving over rounds, though with language models standing in for experts and on general question answering [9]. Rules have been chosen actively for experts to judge in weak supervision [29, 30]. A framework wraps any agent and turns human feedback into rules, for business tasks [31]. Rigour checks have been enforced in code for AI scientists [7], an agent's statistical procedure has been proven correct in Lean [32], reasoning steps have been checked with a solver [8], and notebooks have been linted for statistical errors [33]. Current evaluations of agentic data science are criticised for overlooking methodology and domain knowledge [34]. To our knowledge, nobody selects claims for experts by rule conflicts, writes expert verdicts as logic rules with provenance, checks omics claims on knowledge, data and code in one engine, or uses replication in independent CKD cohorts as the endpoint. Table 2 summarises.

Table 2: Closest work and what it leaves open.

Work	What it does	What is missing here
GRACE [9]	claims checked against an evidence graph; uncertain claims to experts; learning curves	simulated experts; general QA; no statistics or code rules
IWS, Nemo [29, 30]	active selection of rules for experts	label data, not agent claims
Agent Gym [31]	harness around any agent; feedback becomes rules	business agents; no scientific endpoint
OmniScientist [7]	rigour and claim checks enforced in code	no omics rules; no independent cohort
Sargsyan [32]	Lean-verified statistical procedure	does not check claims; no biology
VeriCoT [8]	solver checks of reasoning consistency	logic only
STAT-LINT [33]	statistical errors in analysis notebooks	not single-cell; not agents
GeneAgent [35]	agent claims checked against databases	a lookup, not a declared rule set
BiomniBench [6]	process-level grading with expert rubrics	no independent-data outcome
Johri et al. [36]; OpenScientist [37]	agent conclusions scored on an independent cohort	cancer and myeloma; no process checks
Nakatsuka et al. [23]; Garg et al. [38]	cross-study replication of single-cell findings	brain disease and COVID-19, not kidney

B The rule engine in detail

Table 3: Main rules, by layer. Hard rules block a claim; soft rules flag it for expert review.

Rule	Layer	Type	If broken
statistics per person: aggregate per person before comparing, or use a mixed model	analysis code	hard	wrong
p-values comparing clusters found on the same data are descriptive only	analysis code	hard	wrong
many tests require a multiple-testing correction	analysis code	hard	wrong
every reported number is traced to code output	analysis code	hard	wrong
the test cohort contains nobody from the discovery cohort	data	hard	wrong
enough people per group	data	soft	flag
association evidence is not stated as “causes”	domain knowledge	hard	wrong
within one report, a gene in one cell type does not go both up and down	domain knowledge	hard	wrong
genes and cell types exist (HGNC, Cell Ontology)	domain knowledge	hard	wrong
marker genes match the cell type (HuBMAP kidney table)	domain knowledge	soft	flag
the direction agrees with known kidney injury states	domain knowledge	soft	flag
a gene called a CKD target has genetic support	domain knowledge	soft	flag

The rule base has three families. *Statistics* rules require per-person aggregation or a mixed model [13, 39, 40]; mark p-values for markers between clusters defined on the same counts as descriptive unless recomputed with a valid method such as count splitting [41, 42]; require multiple-testing correction [33]; and require a minimum number of people per group. *Data and logic* rules require every reported number to be traced to code output, no overlap between discovery and test cohorts, association evidence not to be stated as causation, and internal consistency of directions within a report. *Knowledge* rules require genes (HGNC) and cell types (Cell Ontology) to exist, and, as soft rules, marker genes to match the cell type according to the HuBMAP kidney table v1.8, which lists 67 cell types and 186 marker genes under CC BY 4.0 [43, 44], directions of change to agree with known injury states [18], and CKD targets to have genetic support [45]. Knowledge rules carry versions and allow exceptions because kidney knowledge changes. Rules catch errors that sound wrong; they do not catch biology that sounds right but is wrong, nor confounding, which is why experts and independent data follow. How many errors of each kind they catch is an estimate until measured.

Tools. The rules run on a logic tool rather than a theorem prover such as Lean, which is built for mathematical proof and decides only true or false, while most biological knowledge holds “usually”. pySHACL checks the record format and identifiers [46]; the ELK reasoner precomputes Cell Ontology and UBERON relations; the answer set programming system clingo runs hard rules as constraints and soft rules as weighted constraints [14], and clingo-explaid returns the smallest set of conflicting rules, which is what the clinicians see [15]. ProbLog can optionally combine several lines of evidence into a plausibility score [47]. Answer set programming has a precedent in biology: checking whether large regulatory networks agree with expression data and reporting minimal conflicting sets [48, 49]. There the input was experimental data; here it is an agent’s claims, but the task has the same structure.

C Experts in the loop

Selection. Flagged claims are ranked by two signals: the size of their rule conflicts and the uncertainty of the soft rules involved. The selection is compared with random selection, with selection by the agent's own confidence, and with the selection rules of earlier work [9, 29]; the primary measure is errors caught per expert-minute.

Verdicts. Experts receive one claim at a time, with the soft rules it breaks and the code and output that produced it, and give one of three verdicts: real error, rule does not apply (possibly new), or unsure. The experts are real people, at least one nephrologist; one reviews all selected claims and a second about 20% of them, so agreement can be reported.

From verdicts to rules. A real error can become a new hard or soft rule; "rule does not apply" lowers the weight of that soft rule. Each new rule records the expert, claim and round it came from, and is switched off if it keeps raising false alarms in later rounds [17]. Rules learned with one agent and cohort are tested on another, to see whether they transfer.

Rounds and risks. Within each round the rule base is frozen; between rounds it is updated, and progress is measured on data not used to update the rules, so the harness does not grade itself. Expert criteria drift while experts grade [50], so verdicts are recorded per round and checked for consistency. An agent could learn to satisfy the rules' surface; reading code and data rather than declared fields limits this.

D Integration around agents

The harness does not change the agent. The agent under test can be a biomedical agent such as Biomni [51], a general coding agent, or NephroAIX if its code can be obtained [1]; general coding agents often score best and biomedical specialist agents show no consistent advantage [6, 52]. The agent runs its code in a controlled environment, the harness extracts claims from the output, and the agent may not declare a task complete until the engine lets its claims through. Other components can be switched on and off for ablation; Table 4 summarises their evidence. Whether such components make agents more reliable is not settled: a harness change can move a model more than a new model generation [6], but on command-line tasks model choice usually mattered more [53]; we report what is measured.

Table 4: Ablatable harness components. Numbers are from the cited abstracts unless stated.

Component	What it does	Evidence
Independent verifier	re-derives numbers, reruns tests, checks citations; the agent may not finish until it passes	self-correction without external signal hardly helps [54]; tool-based and database checks help [35, 55]
Kidney skills	curated KPMP knowledge, cell states, workflows; controls: none, self-written	curated skills raised the average pass rate from 33.9% to 50.5% [56]
Statistical guardrails	per-person statistics; donor-overlap checks	per-cell tests inflate false positives [13]; not yet measured in agents
Number checker	every number traced to code output	a checker with rules fixed in advance reached a precision above 90% (78 of 85) [57]
Claim ledger	typed records; context assembled per turn	longer input alone cut performance by 13.9–85% [58]; a state-externalising harness gained 11.4 points on retrieval [59]; design idea from [60]
Probabilities and abstention	claims carry probabilities; "cannot be tested"	the best of 17 models got both sides of paired act/abstain tasks right 59.5% of the time [61]
Recurrence filter	keep hypotheses that recur in m of k runs	capability gains brought little reliability gain [62]
Sub-agents	analysis, literature and verification split; verifier fixed	multi-agent setups ranged from $+80.8\%$ to -70.0% against a single agent [63]

For context, the same model gained 9.4 to 13.5 points from a harness change on BiomniBench, against 3.8 points for a model generation, and rerun standard deviations per task were 7.3 to 11.5 points [6].

E How the evaluation data are built

The evidence report card. Suppose an agent states that "gene X is a good CKD target; it is raised in injured proximal tubule cells". The sentence cannot be marked right or wrong as it stands, but it splits into predictions that can be

checked in data, each against independent evidence (Table 5). Each prediction receives one of four results, *holds*, *does not hold*, *underpowered* or *cannot be tested*. When an agent proposes a drug, what we check is the premise behind it.

Table 5: The evidence report card for the example hypothesis "gene X is a good CKD target; it is raised in injured proximal tubule cells".

Prediction	Data	How	Strength
X is raised in proximal tubule cells of CKD patients	cohorts from other groups, never seen by the agent	aggregate per patient; compare CKD with controls	medium (association)
Higher X, worse kidney function	clinical data of the same cohorts (eGFR, fibrosis)	correlation with eGFR	medium (association)
X is genetically linked to CKD	CKD genome-wide association studies	is X near a CKD risk locus?	stronger (close to causal)
An effective drug brings X down	biopsies before and after a drug; treated animals	compare treated with control groups	stronger (intervention)

A paper's data, not its conclusions. Everything we use was released with published papers, but a paper holds two different things. Its *data* are the released measurements, for example each cell's gene expression and which patient it came from; its *conclusions* are what the authors wrote. The data are the main source of answers: we recompute with our own code and one set of rules, which keeps comparisons fair and is the only way to check a new conclusion that no paper contains. Published conclusions are never answers; they are used only to pick the findings for the reference line and to test whether an agent repeats the literature. Results published after a model's training cutoff can be used directly, since the model cannot have read them. As an illustration (not yet computed), Abedini et al. [19] report that injured proximal tubule cells expressing HAVCR1 (KIM-1) increase in CKD, and released per-cell data with patient and disease labels. If an agent analysing Lake et al. [18] claims that HAVCR1 rises in the proximal tubule in CKD, we aggregate each patient's proximal tubule cells in the Abedini data, compare CKD with controls and apply the pre-registered rule; if it names a gene the paper never mentions, the data can still be computed. Our result will probably agree with Abedini's conclusion, so an agent may be right by recall rather than analysis. Three controls separate the two: a literature-only model that never touches data as a baseline, questions with shuffled patient labels, and answers the agent cannot have read (post-cutoff results and, later, possibly unpublished data).

Removing shared donors. The same person must never be both in the data given to the agent and in the data used for checking. We list every donor identifier in every dataset and remove overlaps. The overlaps we found are real (Appendix F): Abedini and Dumoulin are held out together and never test each other, and the Wilson controls are excluded wherever the Muto donors appear. The discovery cohort of Lake et al. [18] uses a different identifier scheme, so overlap can be neither shown nor excluded by identifier; given its different tissue sources we consider it unlikely. Integrated releases that merge several cohorts [64] cannot serve as held-out data.

One format, rule fixed in advance. Statistics are per person, not per cell: within each person we sum the expression of one cell type (pseudobulk) and then compare patients with controls, because per-cell tests produce many false positives [13]. We compare genes and directions, not cell-type names, because cohorts name the same states differently. The decision rule is written down before results are seen (pre-registration), and "does not hold" is kept apart from "underpowered": a small cohort that shows no difference refutes nothing.

The reference line. An agent's score is compared with that of people. We take a set of published CKD findings and test them in independent cohorts under exactly the same rule; after shuffling patient labels, almost none should hold, which also tests the yardstick. The most cited finding, more injured proximal tubule cells in CKD, must be defined by genes rather than names because cohorts define the state differently [18–20]. The link between fibrotic or immune microenvironments and kidney function can be tested with spatial data [19, 28]. The origin of myofibroblasts reported by Kuppe et al. [21] is unsuitable, because the marker used to sort the cells is part of the finding. Cross-study replication has been measured for neurodegenerative disease and COVID-19 [23, 38]; we therefore claim, to our knowledge, the first systematic test in kidney, not the first test of single-cell replication. The reference line is a result in its own right.

Questions. There are three main kinds of question. Target questions ("which genes are worth pursuing as CKD targets?") are scored row by row on the report card; mechanism questions ("what happens to the proximal tubule in CKD?") mainly by replication in independent cohorts; drug questions ("which cells does an SGLT2 inhibitor act on?") against treated groups. Two kinds of control question catch recitation and the failure to say "I don't know": shuffled-label questions, whose correct answer is "no effect", and untestable questions that ask for information the data

do not contain, such as one person's change in eGFR. Optional artifact questions inject known technical artifacts into one group to test whether an agent takes them for disease or a drug effect. A pilot uses about 20 questions; the full benchmark about 50 to 60, comparable to related benchmarks, and the true limit is what the independent cohorts can answer. Question lists, the published findings for the reference line and the soft rules are each reviewed independently by two domain colleagues, with disagreements resolved and agreement reported.

F Datasets

Table 6 lists the public data and their role. People counts, downloads and licences were checked against GEO, Zenodo and the papers on 2026-11-02.

Table 6: Datasets and their role. "Held out" data are never given to the agent.

Dataset	Species	Assay	Role and main caveat
Lake et al. [18]	human	sc, sn, spatial	discovery cohort and planted errors; different identifier scheme; merged releases contain its donors
Abedini et al. [19]	human	sc, sn, ATAC, spatial	test cohort; shares patients with Dumoulin; no aTAL in the public annotation
Wilson et al. [20]	human	snRNA, snATAC	test cohort; controls are Muto donors; 5–7 DKD donors
Kuppe et al. [21]	human	sorted sc	test cohort (eGFR comparison); sorted cells; no healthy controls
Dumoulin et al. [28]	human	CosMx, Xenium	spatial evidence; about half DKD; shares patients with Abedini
Bjornstad et al. [65]	human	sc	treatment (SGLT2 inhibitor vs placebo); group comparison only; 11 follow-up biopsies
Wu et al. [66]	mouse	sn	treatment; five treatments, two time points
Abedini et al. [67]	rat	sn, ATAC	treatment (finerenone, spironolactone); 2–5 samples per arm
Balzer et al. [68]	rat	sc	treatment (enalapril); 3 treated rats
Liu et al. [45]	human	GWAS	genetic row; probably known to current models

Caveats. From our check of GEO, Zenodo and CELLxGENE identifiers (2026-11-02): Abedini et al. [19] and Dumoulin et al. [28] share patients (for example identifiers HK2844 and HK2924); the controls of Wilson et al. [20] are Muto et al. [26] donors and part of its diabetic samples come from Wilson et al. [27]; Lake et al. [18] uses KPMP and HuBMAP identifiers that do not match the other schemes; the KPMP-merged release of Abedini and the integrated data of Klötzer et al. [64] contain donors of several cohorts. Wilson et al. [20] report VCAM1$^+$ injured proximal tubule cells as more frequent in diabetic kidney disease (snATAC 0.12 vs 0.03, $p = 0.004$; snRNA 0.06 vs 0.02, $p = 0.03$). The public annotation of Abedini has no aTAL state. Kuppe et al. [21] is internally inconsistent in donor counts (13 patients in the text, 11 and 8 per sorted fraction, 15 kidneys in a figure legend) and has no spatial data. Dumoulin et al. [28] contains about half diabetic kidney disease, with controls, people with diabetes or hypertension only and other kidney diseases; its gene panels cover about 1,000 (CosMx) and 5,100 (Xenium) genes; the immune-niche share correlates with eGFR ($r = -0.59$). The public data of Bjornstad et al. [65] have no participant identifiers, so before–after pairs cannot be matched. The single-nucleus data of Tuttle et al. [69] are available only on application. Wu et al. [66] used female mice with renin AAV and uninephrectomy for two days or two weeks; the rat arms of Abedini et al. [67] have 2 to 5 samples and Balzer et al. [68] three treated rats.

G Deployment on JUPITER

The harness runs with open-weight models on JUPITER at Forschungszentrum Jülich, whose accelerator nodes combine Arm-based Grace CPUs with Hopper GPUs [70]. Open weights keep model versions fixed, so experiments can be rerun unchanged, whereas a commercial API may change underneath; data that may only be processed locally stay on the cluster; and many reruns are affordable. Steps can be assigned by data sensitivity, sending steps that touch restricted data only to local models [60]. We measure rerun consistency against commercial APIs, throughput and compute per claim that holds. JUPITER accelerator nodes carry four NVIDIA GH200 superchips each [70]; because the CPUs are Arm-based, some inference and training frameworks need extra configuration, and time is reserved for the first deployment. The deployment records model version, seed and environment for every run. Fine-tuning a very large mixture-of-experts model into a kidney model is not planned first: in a 3-shot setting, medically adapted models were better than their base models in only 12.1% of comparisons and significantly worse in 38.2% [71], and new general

models appear within months. The evaluation produces data no one else has, process records paired with report cards. Later, these can train a small verifier that predicts which conclusions will fail, or serve as a reinforcement-learning reward for a mid-size open model.

H Timeline and planned papers

Table 7: First phase, November 2026 to April 2027.

Month	Postdoc	Student	Checkpoint
Nov	claim format; data download; patient identifier inventory	reading; clingo; first hard rules	claim format fixed
Dec	removal of shared patients; per-patient aggregation; first open model running on JUPITER	statistics and provenance rules	data processed; model running
Jan	pre-registered decision rule; reference line of published findings; experts booked	hard rules complete; tests with planted errors	reference line; first planted-error test
Feb	about 20 questions; pilot runs	soft rules (kidney markers, injury states); conflict explanations; selection method	pipeline end to end
Mar	full runs; expert rounds 1 and 2	rule updates from expert verdicts	learning curve over two rounds
Apr	analysis; preprint	thesis submitted; rule engine released	preprint; engine public

The main paper is a computer-science paper on the harness and its four evaluations. The rule engine can also appear as a software note, and the replication of published CKD findings across cohorts is a biological by-product. Raw data are not redistributed; code, split files, rules and results are released.