

ARBITER: retrieval-augmented generation for automated PROBAST+AI risk-of-bias assessment of HIV, HBV and HCV prediction models: a proof-of-concept validation study

Werner Leber

w.leber@qmul.ac.uk

Queen Mary University of London

Joshua Steele

University of Oxford

Zhiqiang Huo

Queen Mary University of London

Jianhua Wu

Queen Mary University of London

Jasmina Panovska-Griffiths

The Pandemic Sciences Institute, University of Oxford

Article

Keywords:

Posted Date: October 28th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10756052/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: Competing interest reported. WL reports a grant from Gilead Sciences. WL is a member of the London Joint Working Group for Substance Use and Hepatitis C. The remaining authors declare no competing interests.

Abstract

Clinical prediction models for blood-borne virus infections — HIV, hepatitis B (HBV), and hepatitis C (HCV) — are published widely, yet appraising their quality remains labour-intensive and inconsistently applied. We developed the Automated Risk-of-Bias Inspection Tool for Evidence Review (ARBITER) to characterise studies, separate model-development and model-evaluation pathways, and generate a categorical judgment with verbatim supporting evidence from manuscript text for each PROBAST+AI signalling question. We illustrate the application of ARBITER (using the Google Gemini 2.5 Pro API) to ten prediction model studies for HIV, HBV and HCV, comparing its accuracy against Claude Sonnet 4.6 (via its chat interface) and a human statistical expert; one study (Kerschberger 2021) informed pipeline refinement before assessment of the remaining nine. All three assessors achieved substantial agreement with expert consensus: ARBITER Cohen $\kappa=0.701$ (95%CI 0.650–0.754), Manual Reviewer $\kappa=0.724$ (95%CI 0.674–0.774), and Claude Sonnet 4.6 $\kappa=0.714$ (95%CI 0.617–0.794). Error analysis revealed characteristic, partly complementary patterns: ARBITER drew incorrect conclusions from correctly retrieved evidence; the manual reviewer overlooked briefly-reported detail both automated assessors captured; and Claude Sonnet 4.6 erred on the most demanding statistical questions. This reproducible, auditable approach improves through human feedback and, with further validation, could support evidence synthesis and reduce reviewer burden.

Introduction

Early diagnosis of blood-borne virus infections, including human immunodeficiency virus (HIV), hepatitis B virus (HBV), and hepatitis C virus (HCV), remains an important public health challenge. Globally, an estimated 40 million people live with HIV,(1) 47 million with HCV, and 240 million with HBV.(2) Undiagnosed infection can delay access to effective treatment, increase the risk of disease progression, and contribute to onward transmission [1]. Risk-based testing strategies may support earlier identification by helping clinicians target diagnostic testing toward individuals with an elevated probability of infection [2]. The increasing availability of routinely collected health data, alongside advances in statistical modelling and machine learning, has created new opportunities to develop prediction models for this purpose [3,4]. However, strong reported model performance does not necessarily indicate that a prediction model is methodologically robust, clinically applicable, or ready for implementation.

Despite the rapid growth of prediction-model research, concerns about methodological quality and reproducibility remain widespread. In our previous systematic review of blood-borne virus prediction models, 64 of 71 studies (90%) were judged to be at high risk of bias [4]. Similar concerns recur across the broader literature, where inadequate sample size, suboptimal handling of missing data, overfitting, insufficient validation, and incomplete reporting continue to limit the reliability and clinical applicability [5]. As machine-learning approaches become more common, model development and evaluation grow more complex, making rigorous methodological assessment both more important and more resource-intensive.

PROBAST (Prediction model Risk Of Bias ASsessment Tool) assesses the risk of bias and applicability of prediction-model studies across four domains - Participants, Predictors, Outcome, and Analysis.(3, 4) It is the standard framework for the appraisal of both development and evaluation aspects of prediction models, using a set of signalling questions for each. (3, 4) Its extension, PROBAST+AI, accommodates regression, artificial-intelligence, and machine-learning methods, adding criteria for statistical power, feature selection, and algorithmic fairness.(5) Applying it consistently, however, is demanding: reviewers must integrate information scattered across a manuscript, distinguish model-development from model-evaluation pathways, and interpret issues such as sample-size adequacy, resampling procedures, overfitting, and data leakage – making appraisal time-consuming, resource-intensive, and susceptible to reviewer variability.(6)

Large language models (LLMs) may offer support for this task. Early work applying LLMs to ROBINS-I (Risk Of Bias In Non-randomised Studies – of Interventions) assessments, randomised controlled trials, and complementary-medicine trials shows that structured prompting and human–AI collaboration can match expert reviewers while improving efficiency.(7-9) Yet the same studies show that automated appraisal remains error-prone: LLMs may overlook relevant details, misinterpret retrieved evidence, or apply criteria inconsistently. The practical question is therefore less whether an LLM can reproduce expert judgments than how automated and human assessment should be combined – which tasks each performs reliably, where each fails, and whether the two compensate for one another.(8) This is difficult to answer from agreement metrics alone, which capture how often assessors concur but not where or why they diverge; and existing work has largely addressed intervention research, leaving the more technically demanding appraisal of prediction-model studies comparatively underexplored. Instrument development lags similarly: a recent scoping review of appraisal tools for clinical AI identified only two risk-of-bias tools, both still at protocol stage.(10) PROBAST+AI has recently appeared, but evaluation of its application – both automated or manual – is still pending.(5)

Retrieval-Augmented Generation (RAG) offers a way to make automated appraisal both more reliable and more inspectable, by grounding LLM outputs in task-specific methodological guidance and manuscript-level evidence. Rather than relying on unconstrained model judgment, a RAG-based workflow can retrieve relevant appraisal criteria, distinguish pathway-specific content, and generate traceable assessments supported by verbatim evidence. Conducting the risk-of-bias appraisal for our previous systematic review(11) was labour-intensive, and the prediction models we assessed were frequently reported in ways that made consistent, criterion-by-criterion judgement difficult. That experience motivated us to develop the Automated Risk-of-Bias Inspection Tool for Evidence Review (ARBITER), which performs study characterisation, separates model-development and model-evaluation pathways, and generates methodological assessments with structured supporting evidence for each PROBAST+AI signalling question. In this paper we illustrate the application of ARBITER to prediction models for HIV, HBV and HCV infection and compared its accuracy against a human statistical expert reviewer and a second, independent LLM assessor (accessed via Claude’s Project interface). In the process, we analysed not only how far the three agreed but also contrasted the type and source of each disagreement. This allows us to explore the situations where automated and human appraisal may fail,

whether those failures are complementary i.e. whether a human-in-the-loop could catch the errors the automated pipeline makes, and similarly, would the pipeline catch the errors the reviewer may make. We developed ARBITER as a proof-of-concept and it is not intended to replace methodological expertise but to complement it via supporting scalable, transparent, and auditable human–AI workflows for evidence synthesis and the trustworthy evaluation of clinical prediction models.

Results

Statistical Benchmarking Results

Quality tier assignments were generated by the RAG pipeline (SQ1.0) for all ten validation papers and reviewed by the first author (W.L.); no study type misclassifications were found. Before conducting independent domain assessments, all three assessors received the tier assignments to ensure consistency in assessment standards. Kerschberger 2021(30) was utilised for iterative pipeline refinement and is excluded from the primary validation corpus.

Agreement with Expert Consensus

Across 200 SQ-level observations (nine papers), 99 (49.5%) were concordant across all three assessors, whilst the remaining 101 (50.5%) contained a discrepancy: Minor in 31 (15.5%), Substantive in 55 (27.5%), and Major in 15 (7.5%) of the observations (Figure 3, panel A). Pairwise inter-rater agreement was moderate for all three assessor pairs (weighted Cohen's κ with linear weights: ARBITER vs Manual Reviewer $\kappa = 0.503$, 95% CI 0.426–0.577; ARBITER vs Claude $\kappa = 0.513$, 95% CI 0.393–0.600; Manual Reviewer vs Claude $\kappa = 0.551$, 95% CI 0.430–0.651). The Manual Reviewer–Claude pair showed the highest pairwise agreement while the two pairings involving ARBITER were slightly lower, though confidence intervals overlapped substantially (Table 2). Overall, three-assessor Fleiss' κ was 0.496 (95% CI 0.398–0.578), consistent with the pairwise estimates. All three assessors achieved substantial agreement with the consensus judgment: ARBITER $\kappa = 0.701$ (95% CI 0.650–0.754), Manual Reviewer $\kappa = 0.724$ (95% CI 0.674–0.774), and Claude Sonnet 4.6 $\kappa = 0.714$ (95% CI 0.617–0.794), with confidence intervals overlapping across all three assessors (Table 2).

Domain-level agreement varied. Directional three-way agreement was highest in the Participants and Outcome domains (81.5% and 72.2%), intermediate in Predictors (58.3%), and lowest in Analysis (48.5%). Specifically, there were 65 of the 101 discrepancies in the Analysis and 18 of the 36 discrepancies in the Predictors domains (Table 3). In the Analysis domain, disagreements centred on the sample-size adequacy thresholds and on whether conditional signalling questions, such as class-imbalance handling, were triggered. Domain-level Fleiss' κ ranged from 0.014 (Outcome) to 0.439 (Analysis); the two lowest values, Outcome (0.014) and Predictors (0.161), both arose where judgments were predominantly positive (~92% in each) — a distribution that attenuates chance-corrected agreement irrespective of raw agreement — and should be read alongside the exact-agreement rates

rather than on their own. In the Predictors domain the discrepancies were predominantly ARBITER reasoning failures, in which evidence about the model's outputs or analysis was applied to questions about how the input predictors were measured (detailed in Error Analysis).

Discrepancy was elevated in the Analysis domain (64.4%, 65/101 SQs) relative to Participants (25.9%, 7/27), Outcome (30.6%, 11/36), and Predictors (50.0%, 18/36). By pathway, the Development pathway produced 67 discrepant observations across all four domains (46.5%, 67/144) and the Evaluation pathway 34 in the Analysis domain alone (60.7%, 34/56) (Figure 3, panel B). Because the Evaluation pathway comprised only Analysis-domain questions - the domain where disagreement was concentrated - its higher rate largely reflects this composition, with three Evaluation-pathway Analysis questions being the principal contributors: Evaluation SQ 4.4 (8 of 8 evaluation papers), where assessors disagreed on whether the class-imbalance correction trigger condition was met; Evaluation SQ 4.2 (7 of 8), where guidance ambiguity around sample-size adequacy for bootstrap validation produced divergent judgments; and Evaluation SQ 4.6 (6 of 8), concerning whether all model-development steps were replicated within the resampling loop.

A sensitivity analysis, which retains the five-point ordinal scale (rather than collapsing to directional classes), lowered agreement for the two ARBITER pairings (ARBITER vs Manual $\kappa=0.402$, 95% CI 0.332–0.478, $\Delta -0.10$; ARBITER vs Claude $\kappa=0.417$, 95% CI 0.306–0.500, $\Delta -0.10$) while leaving the Manual vs Claude pairing essentially unchanged ($\kappa=0.574$, 95% CI 0.484–0.654, $\Delta +0.02$).⁽¹²⁾ This indicates that part of ARBITER's agreement in the main analysis reflected directional concordance with differing confidence, whereas agreement between the two human/LLM reviewers was robust to confidence calibration. (Table 2).

- Insert Table 2

Two domain-level Cohen's κ values warrant interpretation, as they are low for opposite reasons. In Outcome, the near-zero κ is a distributional artefact: with judgments almost entirely positive, there is little variance for a chance-corrected statistic to capture, and it does not reflect poor agreement. Analysis, by contrast, returned the highest domain κ despite the lowest raw agreement, with the most evenly distributed judgments; its disagreements were largely defensible, reflecting differing strictness on sample-size thresholds, ambiguity over trigger conditions, and genuine expert differences on difficult questions.

- Insert Table 3

Failure Type Analysis

Failure type analysis across 101 discrepant SQ-level observations identified 104 failure type components after parsing compound entries (where two distinct failure modes were independently present across different assessors) (Figure 3, panel C). Reasoning failure was the predominant cause (39 components, 37.5%), followed by Retrieval failure (17, 16.3%), Criterion ambiguity (17, 16.3%), Genuine disagreement

(15, 14.4%), Threshold miscalibration (15, 14.4%), and a single Judgment error (1, 1.0%). No Classification cascade failures were present in the final corpus following iterative refinement of the Kerschberger 2021 case study.(13)

Reasoning failures were most frequent in ARBITER (21 components), which consistently located relevant evidence but drew incorrect conclusions from it – for example, correctly retrieving variable-level deletion statements but incorrectly mapping them as participant-level missing data handling (Development SQ 4.3, Tang 2018; Wahome 2013) (See the workbook file ARBITER Aggregate Statistics in Supplementary Materials). The manual reviewer's retrieval failures (5 components) were concentrated in a single recurrent pattern: on Development SQ 2.2, brief pre-processing descriptions embedded mid-Methods were missed across four papers (Balzer 2020; DeWegheleire 2021; Wahome 2013; Yin 2018) despite being correctly identified by both automated assessors. Claude Sonnet 4.6 showed the lowest overall error count, but its failures were concentrated in the most technically demanding signalling questions: six of seven reasoning failures involved Development and Evaluation SQs 4.1, 4.2, and 4.5, all concerning sample size adequacy, overfitting mitigation, and data leakage. The characteristic error was imprecise application of the underlying statistical constructs – calculating EPV against the final model predictor set rather than the candidate pool (Development SQ 4.1, Peebles 2020), treating documentation of a split-sample procedure as sufficient evidence of leakage avoidance without evidence of how leakage was prevented (Evaluation SQ 4.5, Sharma 2009), and conflating post-hoc validation performance with active overfitting mitigation during model development (Development SQ 4.5, Wand 2012). On Development SQ 3.1 (Wand 2012), both the manual reviewer and Claude Sonnet 4.6 accepted antibody-only seroprevalence as an adequate outcome for a model predicting active HCV infection, whereas ARBITER required RNA confirmation; the discrepancy was resolved as No at final adjudication (W.L.).

The 15 genuine disagreement components represented defensible expert-level differences where no assessor could be identified as clearly wrong, concentrated in the Analysis domain.

Although most of ARBITER's reasoning failures fell in the Analysis domain, its six Predictors reasoning failures were distinctive in kind, concentrated on SQ 2.1 (Balzer 2020; Tang 2018) and SQ 2.3 (Balzer 2020; Peebles 2020; Sharma 2009): evidence about the model's outputs or analysis was imported into a question about how the input predictors were measured. On SQ 2.1 (were predictors defined and assessed similarly for all participants?) in Balzer 2020, ARBITER answered "No" by reading the study's high-risk classification cutoffs – varied by sensitivity target and set separately by region, and reported as the "rate of positive predictions" – as inconsistent predictor definition, although those cutoffs govern only how the model output is dichotomised and the predictors were self-reported identically for all participants (consensus: Probably Yes). On SQ 2.3 (were predictors assessed without knowledge of the outcome?), it answered "No" for Balzer 2020 because candidate predictors were screened "based on univariate associations with the outcome", and for Peebles 2020 because data-driven cut-point selection was treated as a predictor-blinding failure – in both, an analysis-stage use of the outcome rather than a property of how the predictors themselves were measured (consensus: Yes). These ARBITER reasoning failures were of comparable frequency to, but fell on different signalling questions than, the manual

reviewer's SQ 2.2 retrieval failures in the same domain (above); within Predictors the two assessors erred about equally often but in different ways.

- Insert Figure 3

ARBITER Impact following iterative refinement (Kerschberger 2021)

Refinement on the first validation paper addressed discrepancies across all four PROBAST domains — Participants (SQ 1.1, SQ 1.3), Predictors (SQ 2.2), Outcome (SQ 3.3), and Analysis (SQ 4.1, SQ 4.3) — comprising both of the paper's Major discrepancies (SQ 3.3 and SQ 4.3) and four Substantive discrepancies (SQ 1.1, 1.3, 2.2, and 4.1).

These were corrected through three mechanisms (Table 4). A change to the tier-classification prompt — adding recognition of purpose-designed prospective studies — corrected an over-strict Tier 3 assignment that should have been Tier 2, resolving the clustered Participants and Predictors discrepancies in a single upstream fix. Two criteria-prompt revisions resolved the Major discrepancies (for example, clarifying for SQ 3.3 the distinction between eligibility screening and outcome-assessment contamination). And changes to the hard-coded retrieval layer — for example, adding facility-based keywords and replacing single-word with multi-word search phrases — closed retrieval gaps on SQ 1.1 and SQ 4.1.

After refinement, ARBITER agreed with the expert consensus on every previously discrepant signalling question except two that no pipeline change could resolve (SQ 4.2 and SQ 4.4). These question-level improvements did not change any domain-level risk-of-bias judgment: the Analysis domain remained high risk (anchored by SQ 4.1, rated No both before and after refinement), so the paper's overall appraisal was unchanged.

Discussion

In this proof-of-concept study we introduce the Automated Risk-of-Bias Inspection Tool for Evidence Review (ARBITER) tool as an automated retrieval-augmented pipeline for PROBAST+AI appraisal. We also illustrate its application in evaluating clinical prediction models for BBVs by comparing its accuracy against a human expert reviewer (J.S.) and Claude Sonnet 4.6 as independent assessors across ten studies on prediction models for HIV, hepatitis B and hepatitis C. Within our set up, nine of the studies formed the primary analysis, whereas an additional one was used to refine the pipeline. All three assessors reached substantial and generally similar agreement with the expert consensus (Cohen κ = {0.701, 0.724, 0.714} respectively, with overlapping confidence intervals, yet each exhibited distinct types of failures.

LLMs have shown varied outcomes when used for risk-of-bias evaluation. Structured-prompting, embedding the appraisal tool directly in the prompt and providing trial text without retrieval, exhibit a

wide range of performance levels: GPT-4's application to ROBINS-I resulted in only moderate concordance with reviewers (61% raw agreement; Kendall coefficient 0.35),(7) while ChatGPT and Claude, when evaluating randomized trials with a modified Cochrane tool, achieved accuracy rates exceeding 84%.(8) Conversely, breaking down the process into individual signalling questions and combining them using RoB2 decision rules – a structure similar to ours – yielded weak performance (F1 scores of 0.10–0.20) with earlier-generation models, (14) and even a recent reasoning model (ChatGPT-o3), using a structured RoB2 prompt adapted from the same template,(8) achieved only fair agreement with a masked expert panel (median κ 0.33).(15) Retrieval-based systems, which provide the model with evidence passages retrieved from within the trial report, show more consistent performance but still fall short of replacing human judgment: ROBoto2, applying ROB2, achieved a best micro-averaged F1 score of 0.71 against expert manual assessments, its authors recommending that humans remain in the loop, (16) whereas NotebookLM, a general-purpose retrieval tool, showed poor agreement with expert Newcastle–Ottawa scores (intraclass correlation 0.08–0.27), leading its authors to advise against its use.(17) These figures should be considered in light of the inherent limitations of expert consensus: when four seasoned human reviewers used RoB2 without prior calibration, the inter-rater reliability for their overall assessment was low (Fleiss' κ 0.16), and it only reached a moderate level (0.42) when implementation instructions were provided.(18, 19) These statistics are not directly comparable, reflecting different tools, study types and metrics. In summary, variability among raters is a fundamental aspect of risk-of-bias evaluations, impacting both expert human reviewers and automated systems alike. As appraisal tools grow in complexity, reported agreement diminishes, and this reduction is not entirely compensated by more sophisticated models. However, structured guidance enhances consistency for both humans and LLMs. Within this context, ARBITER – a code-based pipeline operating on a local computer connected to an external LLM API, which relies on retrieved methodological and clinical guidance rather than solely on manuscript text - achieved a level of agreement with expert consensus similar to that of an expert human reviewer and a cutting-edge LLM using PROBAST+AI criteria and assessment instructions, while delivering transparent, auditable, and evidence-linked evaluations. These results indicate that future improvements will stem not merely from more powerful models, but from the way models are directed: through structured criteria, retrieved methodological knowledge, and expert reviewers who refine the system with each assessment cycle.

There are two significant omissions in this published work. None of the methods ground evaluation in a curated, domain-specific methodological knowledge base; in the systems reported to date, where retrieval is employed, it is limited to the article under assessment. Additionally, they focus on intervention or observational studies rather than clinical prediction models, which require distinct pathways for development and evaluation, as well as assessments, such as sample size adequacy, overfitting, and data leakage. ARBITER addresses both issues: very few, if any, purpose-built systems base risk-of-bias evaluation on external curated knowledge - such as the PROBAST+AI criteria combined with disease-specific clinical and statistical knowledge - instead of relying solely on the source document and specifically target prediction-model studies.

In this study, we demonstrated that the level of agreement between ARBITER, the human reviewer, and Claude Sonnet 4.6 was comparable, both in pairwise comparisons and against the expert consensus (Cohen κ versus consensus: ARBITER 0.701, Manual Reviewer 0.724, Claude Sonnet 4.6 0.714), although the reasons for this agreement varied. ARBITER's mistakes were mainly due to reasoning errors; the manual reviewer's unique shortcoming was overlooking briefly mentioned study details that both automated assessors identified; and Claude Sonnet 4.6's errors were due to inaccurate application of statistical concepts. Refining ARBITER's analysis on a single paper enhanced its performance, and these improvements were applied to the remaining nine papers.

We developed ARBITER as a retrieval-augmented system to overcome a shortcoming of general-purpose models used for this task: whether they are prompted with the guidance and source text directly, or rely solely on passages retrieved from the study, they tend to produce fluent assessments that lack methodological support. Instead of relying on parametric memory, each judgment is grounded in curated, retrievable knowledge, which also makes ARBITER's outputs verifiable, as each judgment can be traced back to the criteria file and the verbatim evidence it is based on.

Our failure analysis is consistent with this reasoning. The accuracy of the model on a specific fact correlates with how well that fact is represented in the pre-training data. The specialized methodological and diagnostic knowledge required for PROBAST+AI appraisal is part of the long tail of that distribution - it is sparsely represented and not reliably encoded, a gap that larger models only slowly bridge and that curated retrieval directly addresses. ARBITER's reasoning failures clustered precisely where the criteria files did not fully encode the relevant knowledge: in the Analysis domain, and on Predictors questions where evidence about a model's outputs was read as a property of how its input predictors were measured. Because such failures are systematic, they may be amenable to targeted criteria-file revision — and they are exactly the kind of slip a human reviewer is well suited to catch.

Where grounding was adequate, reasoning proofed effective. In the context of Development SQ 3.1 from Wand 2012,(20) both the manual reviewer and Claude Sonnet 4.6 accepted that antibody-only seroprevalence was a valid outcome for a model predicting active HCV infection. However, ARBITER uniquely demanded RNA confirmation to differentiate between active and past or resolved infections, which is the methodologically correct standard and was upheld during adjudication. Although this is just one instance, it illustrates the opposite of ARBITER's typical shortcoming: when the criteria files contained the correct disease-specific standard, the grounded pipeline reasoned accurately on a point that both other evaluators had — incorrectly - agreed upon. In this scenario, where the criteria files included the correct standard, ARBITER retrieved and reasoned accurately, like how human experts can improve as they gain more domain knowledge.

The feedback-and-adjudication loop is fundamental rather than supplementary. Each cycle of human review identifies where grounding is incomplete enabling updates to criteria files or code, thereby enhancing grounding and reasoning over time. This process also redefines the optimal integration of automated and human evaluation: our assessors exhibited distinct, partially complementary

weaknesses, allowing them to compensate for each other. ARBITER's role is not to replace expert evaluation but to serve as a transparent, improvable secondary reviewer — one that consistently applies grounded criteria, highlights discrepancies for human review, and evolves through such scrutiny. Its multi-layered structure, featuring three independent assessors and a three-person panel (W.L., J.S., Z.H.) adjudicating Substantive and Major discrepancies, reflects established systematic-review practices. Additionally, the explicit taxonomy of failure types transforms disagreements into distinct, analysable categories, distinguishing correctable errors from genuine interpretative differences. In line with these characteristic differences, a sensitivity analysis that maintained the Yes/Probably Yes distinction reduced agreement for the two ARBITER pairings ($\Delta \approx -0.10$) while keeping the agreement between the manual reviewer and Claude Sonnet 4.6 largely unchanged. This suggests that ARBITER allocates confidence between definite and tentative judgments somewhat differently than the other two assessors, although whether this is due to the underlying model, the prompts and criteria files, or the limited sample size cannot be determined here.

There are several limitations to consider. This study is in its early stages: it relies on nine held-out studies from a single clinical domain of BBV infections, which demonstrates that the study works in principle. The ratings for signalling questions are grouped within studies, and although we used study-level clustered bootstrapping to address this, precision remains limited. Refining one paper does not demonstrate how broadly such improvements can be applied. Future research should validate ARBITER on larger numbers of blood-borne virus prediction models and across other clinical domains to test its generalisability. ARBITER's dynamic prompt assembly for each question is complex, affecting interpretability, even though the criteria files remain constant across papers. Automatically dividing into development and evaluation pathways might limit the model's contextual understanding, especially for studies with poor reporting. Quality-tier assignments were reviewed by the first author without independent verification by all three assessors. Additionally, the text analysed did not include tables and figures; the statistical knowledge base focuses on traditional methods and may not fully represent machine-learning concepts; and sample-size adequacy was assessed using events-per-variable guidance instead of the minimum sample-size criteria outlined by Riley et al.(21)

Future research will look to further validate the effectiveness of ARBITER by applying it to a broader and more varied collection of studies, including the remaining prediction-model studies identified in our systematic review. Additionally, we plan to explore the integration of tables and figures, potentially through multimodal retrieval on the rendered page. Since the system operates independently of any specific model, it can theoretically be executed locally, providing data security control that is less critical for published studies but significantly important for contexts like manuscript peer review. Currently, this is limited by hardware, but we anticipate that this limitation will decrease as more efficient models become available or local machines with higher memory become capable of running more advanced models. This work was out of scope for this paper, but we are planning to explore it further in future. Furthermore, the retrieval-augmented method could be modified for use in other clinical areas and study types, such as RoB2 or ROBINS-I for intervention studies. However, this would necessitate not only replacing the criteria files but also altering the prediction-model-specific elements of the pipeline, such

as pathway separation, which are not applicable to trials. This is again out of scope here, but an interesting possible future extension of our work.

This study introduces the Automated Risk-of-Bias Inspection Tool for Evidence Review (ARBITER) as an evidence retrieval-augmented generation system can appraise and summarise published prediction-model studies. Applied to HIV, HBV and HCV in this proof-of-concept study and tested against human statistician and an LLM, ARBITER grounds each assessment in a curated knowledge base and files containing disease-specific criteria. For each signalling question, it creates a unique prompt that combines text from the paper with the relevant criteria and retrieved statistical knowledge. ARBITER provides a clear, reproducible, and adaptable way to assess risk of bias, aiming to assist expert reviewers rather than replace them.

Methods

System Architecture and Overview

We developed an Automated Risk-of-Bias Inspection Tool for Evidence Review (ARBITER), as an automated system to evaluate clinical prediction models and here illustrates its application to HIV, HBV, and HCV (Figure 1). The system performs two complementary analyses: PICOTS characterisation describing what the study investigated, and PROBAST assessment across its methodological domains evaluating how the study was conducted. PROBAST+AI extensions provide specific guidance for assessing models incorporating artificial intelligence and machine learning techniques.

- Insert Figure 1

ARBITER is created as follows. Firstly, we implemented, in Python, a Retrieval-Augmented Generation (RAG) tool combined with a Model Context Protocol (MCP) server for calculation of the events per variable (EPV) ratio, an estimate for sample size in prediction models. Then, via an application programming interface (API), an LLM interprets each study text to extract PICOTS elements and applies the PROBAST criteria, with assessment rigor adjusted to the study quality. For studies with both development and validation phases, we characterised and assessed each phase separately using the pathway-specific criteria.

We created ARBITER to be a disease-agnostic information retrieval tool with the PROBAST+AI criteria, pathway separation, retrieval design, and reporting remaining as general as possible. In this proof-of-concept study we illustrated and validated ARBITER's application to prediction models for HIV, HBV, and HCV, via embedding criteria prompts, risk factors and clinical examples to be BBV-specific. Adapting the pipeline to another clinical domain therefore requires changes only to these prompt and code components, not to the underlying architecture. We refer to the general, domain- and model-independent core of the RAG as the architecture, and to the BBV elements layered onto it – the embedded clinical examples, laboratory methods, risk factors, and search terms – as the configuration.

ARBITER System Components

Large Language Model (LLM) Interface

ARBITER is agnostic to the underlying LLM: the pipeline can run either through a local API, with the model hosted on local hardware, or through a provider-hosted cloud API. The system itself — its code and prompt library — was developed primarily with Claude Sonnet 3.7–4.5 (Anthropic) as a coding and authoring assistant, with some assistance from Gemini 2.5 Pro (Google) for edge cases; these models were used to build the pipeline and are distinct from the model that performs inference when it runs. During development and engineering, the pipeline was run locally on a Qwen3-8B model (Alibaba) served through LM Studio 0.3.30 on a Mac mini (Apple M4, 16 GB RAM), accessed as a local API. For the validation reported here, the locally deployed Qwen3-8B model risked reduced accuracy and impractically long processing times; the pipeline was therefore run on a higher-capacity, provider-hosted cloud API (Gemini 2.5 Pro) and benchmarked against a manual reviewer (J.S.) and Claude Sonnet 4.6 (accessed via its chat interface).

ARBITER Knowledge Base and Retrieval

Prompt Library Organisation

We organised a hierarchical prompt library storing assessment frameworks for all 34 signalling questions across development and evaluation pathways, plus six applicability assessments (Figure 1). Each question has a dedicated criteria file (e.g., `participants_development_q1_1_criteria.txt`) defining its analytical structure — two-pillar frameworks for Participants, three-pillar frameworks for Predictors and Outcome, and the events-per-variable (EPV) framework for Analysis — together with tier-specific decision tables mapping evidence quality to categorical judgments (Yes, Probably Yes, Probably No, No, and a single neutral category combining Unclear, No Information and Not Applicable) under lenient, moderate, or strict standards. These files are fixed, comprehensive, and specific to the BBV context, and are reused unchanged across all papers.

Vector-Indexed Knowledge Base

We indexed a methodological knowledge base in a ChromaDB vector database, comprising the PROBAST and PROBAST+AI guidelines,(4, 5) a statistics textbook,(22) and statistical-methods papers, (23-36) enabling semantic retrieval of the passages most relevant to the model under assessment. Neither the criteria-file prompt library nor this knowledge base is fed to the model as-is. For each signalling question, the pipeline dynamically assembles a unique fresh prompt (Figure 2) from the portions of the relevant criteria file that apply to the specific prediction model, combined with passages semantically retrieved from the knowledge base for that model. Because retrieval is driven by conceptual similarity to the model at hand, it returns only pertinent methodological content and does not draw in

material from unrelated clinical fields (e.g., diabetes or cancer); this keeps each prompt focused and reduces processing time by 30–50%.

- Insert Figure 2

ARBITER Document input and indexing

For each study, the Title, Abstract, Methods and Results were provided as a plain-text file; tables and figures were excluded, as their extraction did not yield technically reliable results during development. Each file was named using an Author–year–DOI convention (e.g. Burns_2022_10.1093/cid/ciac775), which served as the document identifier enabling the retrieval system to locate and target the correct manuscript. ARBITER segmented the supplied text by section and indexed it for downstream retrieval; this structured text formed the sole input to all classification and assessment steps. For the studies appraised here, papers were retrieved and the relevant sections prepared manually prior to ingestion.

ARBITER Prompt Engineering and Optimisation

We generated the criteria files and supporting prompts using the chat interface of Claude Sonnet 3.7–4.5 (Anthropic), refining them through iterative consultation of clinical guidelines and our systematic review, with comparative edge-case testing via the chat interface of Gemini 2.5 Pro (Google). To ground each assessment in real-world clinical knowledge, we integrated disease-specific information with the criteria files: the types of risk factors relevant to each model (e.g. behavioural versus clinical), the study settings in which predictors and outcomes were identified (e.g. primary care, substance-use services, or hospital), and – critically for outcome appraisal – the tests required to distinguish active from resolved or past infection (e.g. HBsAg for hepatitis B and RNA confirmation for hepatitis C). These disease-specific components form the blood-borne-virus framework that is layered onto the otherwise general disease architecture.

ARBITER Assessment Pipeline

PROBAST+AI evaluates prediction models using standardised signalling questions across four risk-of-bias domains: Participants, Predictors, Outcome, Analysis). Each question is answered with a categorical judgment, which informs the risk of bias at both the domain and study levels. Prior to examining these questions, ARBITER applies its own three-step classification system that determines how each paper is routed (development vs. evaluation pathway) and how strictly it is judged. To align with the codebase, these steps maintain the SQ prefix - SQ0.1 (study design categorisation), SQ0.5 (pathway separation) and SQ1.0 (quality tier classification) – and are distinct from the PROBAST signalling questions. For example, SQ1.0 serves as the quality-tier classifier and does not pertain to the Participants-domain signalling questions (1.1–1.3). Initially, each study is categorised by design (SQ0.1), which then guides two analyses of the same paper (Figure 1): a PICOTS characterisation of what the

study investigated, and a PROBAST+AI appraisal of how it was conducted. This approach allows for the same study type to be applied to both the descriptive and appraisal part of the analysis.

Stage 1: Study Design Categorisation (SQ0.1)

The pipeline first each study into one of seven standard designs:

- Development of a novel prediction model
- Development with internal validation
- Development with external validation
- Development with both internal and external validation
- Internal validation of an existing model
- External validation of an existing model
- Updating or extending an existing model

Studies were categorised using a confidence rating generated by the LLM. A rating between 0.90 a0.95 ensures accurate PROBAST pathway assignment. A rating of 0.5 triggers additional Regex pattern matching with weighted scores. Should the rating remain below 0.5, threshold manual researcher verification is required to avoid incorrect pathway assignment. The study type identified guides both the PICOTS characterisation and the PROBAST appraisal pathway, ensuring a consistent study type is assigned to both. The assigned pathway determines the signalling questions used: development-only studies undergo 16 questions (Participants 1.1-1.3, Predictors 2.1-2.4, Outcome 3.1-3.4, Analysis 4.1-4.5), and evaluation-only studies 18 questions (Analysis extended to 4.1-4.7). Since the Participants, Predictors and Outcome domains share the same signalling questions across both pathways, they were assessed once for the development pathway in this study. However, the Analysis domain, which has unique signalling questions for each pathway, was evaluated separately for both the development and the evaluation pathways. A mixed study (development and evaluation) thus underwent the 16 development signalling questions plus the seven evaluation Analysis questions, totaling 23, instead of a full duplicate assessment. SQ0.1 initiates a dependency chain (SQ0.1 → SQ0.5 → SQ1.0); the following steps, which differentiate pathway content and assign quality tiers for the PROBAST assessment, are detailed below.

Stage 2: PICOTS Characterisation

PICOTS characterisation offers a systematic, descriptive summary of what each study investigated - its scope and design – differing from the PROBAST domains, which evaluate the development and evaluation of a model using predefined signalling questions. The LLM characterises each study across eight PICOTS components: Population, Index Model, Comparator, Outcomes, Timing, Setting, Exclusions, and Study Type. The first seven components are extracted by the LLM from the manuscript sections

(abstract, methods, results) using domain-specific prompts and few-shot examples. This process includes post-processing for standardizing model names and deduplication, standardizing outcomes for co-infection studies with multiple BBVs, and categorizing settings into standard classifications. The Study Type component was not re-derived here but linked to the SQ0.1 classification and expanded with descriptive methodological detail, such as "Development with internal validation (10-fold cross-validation) and external validation (temporal validation across the HVTN505 and HPTN061 cohorts)".

Stage 3: Pathway Separation and Quality Tiering

Two further classification steps configure the appraisal for each study, completing the SQ0.1 → SQ0.5 → SQ1.0 dependency chain. Unlike SQ0.1, these steps serve the PROBAST appraisal only: SQ0.5 determines which text each signalling question sees, and SQ1.0 sets how strictly it is judged.

SQ0.5: Pathway Content Detection. In studies with multiple pathways, this step directs the extracted manuscript text into development and evaluation pools - it assigns, rather than extracts. Using the cohort names identified in SQ0.1 (e.g., "HPTN 068 trial" for development, "Zimbabwe site" for evaluation), the LLM determines which cohort is described in each paragraph and assigns the entire paragraph to the corresponding pathway, maintaining the methodological context. Studies with a single pathway do not need this split. The pools specific to this pathway are then forwarded to the domain assessment, where each signalling question draws solely on its pathway. The process for selecting passages specific to each question is described in Analysis at the Level of Individual Questions.

SQ1.0: Quality Tier and Model Purpose. This step establishes the level assessment strictness by assigning a quality tier and determining the model's purpose. The LLM assessed four quality criteria – (1) a formally constituted data source, (2) systematic protocols, (3) formal oversight, and (4) institutional scale, assigning a tier based on these evaluations:

Tier 1 (4/4): lenient standards

Tier 2 (2–3/4): moderate standards

Tier 3 (0–1/4): strict standards

This aligns assessment standards with study quality (e.g., Tier 1 required for EPV ≥ 10 ; Tier 3 required EPV ≥ 20 plus overfitting mitigation). Once assigned, tiers are consistently applied across all domains. The LLM also determines the model's purpose (diagnostic, prognostic, or risk stratification), which influences outcome events included in the EPV calculations. The reasoning is that a study built on a robust, well-governed data source - such as a model derived from a retrospective analysis of randomised-controlled-trial data, with predefined protocols, ethics approval, and multisite conduct - can be considered methodologically sound even if reporting is incomplete, and so is held to more lenient standards. In contrast, a single-site study with a small sample and no protocol lacks these safeguards and must document its methods explicitly, thereby attracting stricter standards.

Stage 4: PROBAST Domain Assessment

The LLM assessed the methodological quality across four PROBAST domains - Participants, Predictors, Outcome, and Analysis - processing each sequentially but independently (Figure 1). Tier assignment determined assessment standards. Pathway-specific content ensured development questions analysed development text while evaluation questions analysed validation text.

Dynamic Prompt Assembly. For each signalling question, a question-specific prompt at run time, referred to as the dynamic prompt, at run time. This prompt consisted of a fixed component - the relevant sections from the criteria files – along with the components selected for each question: semantically retrieved excerpts from the PROBAST guidelines and the manuscript passages specific to that pathway (see Knowledge Base and Retrieval). Two assessment-specific elements were then added: the tier-based standards assigned at SQ1.0 and the required JSON (JavaScript Object Notation) response structure. For example, when assessing Participants Question 1.1 for a Tier 2 development study, the dynamic prompt combined the two-pillar framework and moderate assessment standards (criteria file), retrieved PROBAST guidance on data sources, the development-cohort text, and the JSON instructions to return structured reasoning, a categorical judgment, a justification, and supporting quotes (Figure 2).

Analysis at the Level of Individual Questions. In contrast to PICOTS, where the LLM accesses entire sections, each PROBAST signalling question is provided only with the excerpts pertinent to that specific criterion. This isolation ensures evaluation based exclusively on relevant evidence. For example, Question 2.1 is only given the text related to predictor measurement, while Question 4.1 provided only with sample information. Assessment frameworks differ across domains.⁽⁵⁾ The Participants domain focuses on data provenance and collection (two-pillar structure); Predictors evaluates consistency of definition across all participants, measurement and processing standardisation, and fairness across groups (three-pillar structure); and Outcome assesses definition quality and measurement consistency and integrity across participants (three-pillar structure). Analysis uses an EPV framework quantifying statistical power through the ratio of outcome events to candidate predictors. The purpose of the model determines how criteria are being applied. Diagnostic models require cross-sectional data and concurrent measurement; prognostic models require longitudinal cohorts with temporal separation between predictors and outcomes. Outcome criteria differentiate between identifying existing cases (diagnostic) and new occurrences (prognostic), affecting testing algorithms and follow-up needs.

Statistical Power Assessment. The LLM employed tier-specific EPV thresholds calibrated to model complexity. Traditional regression required 10–20 events per candidate predictor across tiers. Machine learning models required substantially higher EPVs (≥ 50 –200 depending on algorithm complexity) and deep learning models ≥ 500 events, reflecting greater parameter counts and overfitting propensity.⁽⁴⁾ We counted all candidate predictors initially considered, not merely those retained in the model, since variable selection consumes statistical power. Event counts were adjusted for model purpose: prevalent cases (diagnostic), incident cases (prognostic), total population (risk stratification). An MCP server connected to the LLM (Figure 1) processed case counts and predictor numbers.

Risk Aggregation. Risk aggregation follows hierarchical PROBAST methodology (Table S1, Supplementary Materials). Within each domain, if any signalling question receives a "No" or "Probably No" the domain is classified as high risk. If responses are unclear but do not include high-risk answers, the domain is rated as unclear. A low-risk classification is only achieved when all responses are favourable. The overall study risk is based on the highest domain-level rating amongst the domains. For studies with varying domain ratings, the worst pathway rating within each domain is used. Unlike manual reviewers who might adjust a rating with documented reasons, the automated system strictly follows the hierarchical PROBAST rules without exceptions.

Stage 5: Output and Reporting

We generated structured outputs using a dual-storage system: HTML reports created with Jinja2 templates for easy human review at the paper level, an Excel database in CSV format, and records in a Structured Query Language lite (SQLite) database for systematic querying and aggregation. Reports are formatted based on the study type: three-column layouts for mixed studies (showing parallel assessments) or two-column layouts for single-pathway studies. Each report includes nine sections: paper identification with DOI links, structured PICOTS summaries, study classification with pathway context and quality tier, performance metrics (processing time, token usage), an integrated summary table that consolidates all questions and risk assessments, detailed domain evaluations with expandable evidence, ingestion validation metrics, system warnings, and the complete ingested text for verification (See ARBITER report of Scott et al, 2020, in Supplementary Materials).

The database stored metadata on assessments, report paths, and timestamps (ISO 8601), establishing persistent links between papers and evaluations. This method ensured that assessment data remained accessible for systematic review synthesis while providing human-readable reports for individual review and quality assurance.

ARBITER Application to BBVs

We showcase the application of ARBITER on an iteratively engineered pipeline of a set of separate, non-overlapping ten prediction-model studies for BBVs.(13, 20, 37-44) Studies were randomly selected by Claude (Anthropic) from our previously published systematic review of 71 studies(11)using the following criteria: viral type (HIV, HBV, HCV), study type (retrospective, retrospective Registry, retrospective RCT, cross-sectional, prospective), statistical method (logistic regression, machine learning, other), and risk of bias (high, low) as defined in our systematic review.(11) We conducted systematic validation by comparing assessments from three independent assessors — ARBITER (the RAG pipeline, accessed via the Gemini 2.5 Pro cloud API), a manual reviewer (J.S.), and Claude Sonnet 4.6 (Anthropic, via its chat interface) — each evaluating all ten validation papers independently across both Development and Evaluation pathways (for more detail, see the Comparators section below). The papers were randomly selected from our published systematic review, stratified by blood-borne virus (HIV, HBV, or HCV),

statistical model (machine learning, logistic regression, or other), study type (retrospective, prospective, or cross-sectional), and PROBAST risk of bias (high or low) (Table 1), spanning diverse data sources, settings, and publication years (2009–2023).

These issues, highlighted by the manual and Claude Sonnet 4.6 reviewers, led to revisions in the pipeline prompts and, when necessary, the code. After these adjustments, the pipeline was finalized before evaluating the remaining nine papers. Consequently, the primary validation analyses focus on these nine studies, while Kerschberger et al, 2021 is treated as an interim refinement.⁽¹³⁾ The complete validation set of ten studies is presented as a sensitivity analysis, with further details provided in the Results section.

- Insert Table 1

ARBITER Comparators

ARBITER RAG pipeline. The automated evaluation was generated by ARBITER, utilising the Google Gemini 2.5 Pro cloud API. For each signalling question, ARBITER assembled a prompt from the relevant criteria file, excerpts from the PROBAST and PROBAST+AI guideline retrieved semantically from the ChromaDB knowledge base, and the pathway-specific excerpts from the paper being evaluated; Gemini 2.5 Pro then returned a judgment and justification.

ARBITER's architecture is backend-agnostic: the retrieval pipeline, criteria files and prompt assembly do not rely on the specific LLM used, which functions as an interchangeable inference engine. The system was developed using a locally-hosted model run on a basic MAC mini consumer model, confirming that it can operate entirely within a local environment. On this hardware, however, the local model proved operationally unreliable for full validation — generation was slow and outputs were intermittently incomplete or malformed — so the validation reported here was run on the Gemini 2.5 Pro cloud API. This was a deliberate choice: our aim was to evaluate the RAG pipeline under favourable backend conditions, rather than to evaluate any specific LLM.

Manual Reviewer. One manual reviewer (J.S.), a statistician with expertise in prediction model methodology, independently assessed all 10 validation papers using the PROBAST+AI Explanation and Elaboration guidance and standardised PROBAST+AI assessment templates, blinded to the RAG and Claude assessments. Where the statistical interpretation of a signalling question was ambiguous, the manual reviewer (J.S.) consulted a second statistician (J.P.G.) to inform the assessment.

Claude Sonnet 4.6. As a third independent assessor, we deployed Claude Sonnet 4.6 (Anthropic) via the claude.ai Projects interface (a chat interface) to evaluate all 10 validation papers using standardised PROBAST+AI assessment templates and a version-controlled system prompt embedded in the project configuration, which defined tier classification, judgment options, output format, discrepancy classification rules, and the seven-category failure taxonomy. Claude received identical extracted paper

text to the human and ARBITER assessors and was explicitly instructed to assess independently, without reference to the other assessors' judgments.

Following independent assessment of each paper, Claude was provided with the ARBITER and Manual Reviewer judgments and justifications within the same conversation and instructed to generate, for each paper, a draft consensus statement and a failure report. For each signalling question, Claude classified the discrepancy by level, derived a consensus judgment under the pre-specified directional rules described below, gave a 2–3 sentence justification grounded in paper evidence, and assigned a failure type. Because Claude both assessed the papers and operationalised the consensus, the consensus judgment and discrepancy level were fixed by the predefined directional rules rather than by Claude's discretion. Substantive and Major discrepancies were adjudicated by a human expert panel (W.L., J.S., Z.H.), with W.L. holding oversight of the final adjudication decisions and verifying the accuracy of Claude's consensus and failure-type reporting throughout. J.S. served both as the manual assessor and on this panel; this follows the convention that reviewers resolve disagreements through structured discussion, with the independent members (W.L., Z.H.) settling unresolved cases, and is a deliberate design choice rather than a guarantee of full independence.

A discrepancy was classified as None when all three assessors gave the exact same answer; Minor when all answered in the same direction but differed in confidence (e.g., two Probably Yes and one Yes); Substantive when two shared a direction and the third differed (e.g., Probably No, No, and Yes); and Major when all three answered in different directions (e.g., No Information, Yes, and No).

Claude populated a 14-column Excel workbook per paper (separate workbooks for Development and Evaluation pathways) recording each assessor's judgment and justification, the consensus judgment and justification, discrepancy level, failure type, and triage confidence.

ARBITER AGREEMENT METRICS

Consensus judgments were derived according to a pre-specified majority-direction rule. Three directional classes were defined: Positive (Yes, Probably Yes), Negative (No, Probably No), and Neutral (No Information, Not Applicable, Unclear). When two or three assessors shared a directional class, the consensus judgment was drawn from that majority class — specifically, the more conservative of the two agreeing judgments (e.g. if two assessors returned Yes and Probably Yes, the consensus was Probably Yes; if two returned No and Probably No, the consensus was Probably No). No single assessor carried special weight: even where one assessor's judgment was considered more methodologically defensible based on subsequent failure type analysis, the majority direction was binding. We included the Failure Type and Triage Confidence of any discrepancy to provide additional context to a disagreement as this is missed in a quantitative analysis. For Substantive and Major discrepancies, a reasoned consensus judgment was derived by a human expert adjudication panel (WL, JS, ZH) via discussions and paper evidence. These same directional class assignments were applied to all k computations, collapsing the seven-category judgment scale to the three-class coding used throughout the inter-rater agreement analysis.

We analysed agreement among all three assessors: ARBITER, Manual Reviewer, and Claude Sonnet 4.6. We assessed question-level and pairwise assessor agreement using Cohen's weighted κ with linear weights and three directional categories — positive, neutral and negative.

In addition, we performed a sensitivity analysis on the pairwise assessor agreement using Cohen's weighted κ on a five-point ordinal scale (No < Probably No < Neutral < Probably Yes < Yes), in which Unclear, No information and Not Applicable were combined into a single Neutral category.⁽¹²⁾ This retains the confidence distinction between Yes and Probably Yes (and between No and Probably No) that the three-class coding collapses, while avoiding an arbitrary ordinal ranking among the non-directional categories, which do not fall on a Yes-to-No continuum; it is important to determine whether disagreements were primarily directional or confidence-calibration artefacts.

Fleiss' κ was computed across all three assessors as a pooled agreement metric.⁽⁴⁵⁾ Because signalling-question ratings are nested within studies, 95% confidence intervals for all κ statistics were obtained by a study-level clustered bootstrap, resampling the validation studies with replacement (2,000 iterations) and recomputing each statistic within each resample, with resamples leaving a statistic undefined discarded; given the small number of clusters these intervals are approximate. Point estimates were computed on the full data and are unaffected by the choice of resampling units. Agreement statistics were initially computed using Claude Sonnet 4.6 (Anthropic) and independently verified by a statistician co-author (J.S.). Observer agreements were classified using the method by Landis and Koch, 1977.⁽⁴⁶⁾

For each discrepant signalling question, we assigned a Failure Type from a seven-category taxonomy: (1) Retrieval failure — the correct answer was present in the paper but was not included in the retrieved context; (2) Reasoning failure — relevant content was retrieved but the assessor drew an incorrect conclusion; (3) Criterion ambiguity — the PROBAST+AI guidance was genuinely unclear for the case type, making different interpretations defensible; (4) Threshold miscalibration — the assessor was directionally correct but applied the wrong strictness level for the assigned quality tier; (5) Genuine disagreement — a defensible interpretive difference where neither position was clearly wrong; (6) Classification cascade — an upstream study design misclassification (SQ0.1) propagating incorrect assessment standards to downstream signalling questions; and (7) Judgment error — an assessor selected a response inconsistent with its own stated reasoning or with the model type (for example, applying a prognostic-model framing to a cross-sectional diagnostic model). Compound failure types were recorded where two distinct failure modes were independently present across different assessors.

ITERATIVE REFINEMENT

Following the interim results, ARBITER was refined, as prompted by the first validation paper (Kerschberger 2021)⁽¹³⁾ - the first paper assessed - the first author (W.L.) reviewed all Substantive and Major discrepancies identified by Claude's consensus procedure and revised the RAG system accordingly. Actionable failures were addressed through one of two pipeline modification mechanisms: targeted modifications to domain-specific criteria prompt files (for Reasoning failures and Threshold

miscalibration), or updates to the pipeline codebase (for Retrieval failures and Classification cascade errors). Criterion ambiguity findings were marked for upstream guidance review rather than pipeline modification, as these might indicate ambiguity in the PROBAST+AI guideline wording rather than correctable errors in the assessment system; for example, genuine disagreement findings were recorded as irresolvable and required no pipeline intervention. The refined pipeline was then frozen before the remaining nine papers were assessed (see Validation Study Design), with the manual reviewer and Claude Sonnet 4.6 assessments remaining unchanged.

Ethics approval and informed consent

This study analysed previously published studies only and did not involve human participants; ethics approval and informed consent were not applicable.

Declarations

Acknowledgments

None.

Author contributions

CONTRIBUTORSHIP STATEMENT (CRediT taxonomy)

CRedit Role	Contributors (initials)
Conceptualisation	WL
Methodology	WL, ZH, JS
Software	WL
Investigation	WL, JS, ZH
Data curation	WL
Formal analysis	WL, JS
Validation	WL, JS, ZH, JPG
Writing – original draft	WL
Writing – review and editing	All authors
Visualisation	WL
Supervision to JS	JPG
Project administration	WL

Data Availability Statement

The ARBITER source code, including a demonstration video, was made available to the editor and the peer reviewer during manuscript review. The codebase is available from the corresponding author (W.L.) upon reasonable request.

Funding sources

W.L. is supported by a National Institute for Health and Care Research (NIHR) School for Primary Care Research (SPCR) grant (number FR4 583), under which this work was conducted as exploratory research building on the associated systematic review; some large language model subscription costs were met from the grant. The study received no separate dedicated funding. ZH and JW were supported by Barts Charity (MGU0504). JPG's and JS's work is supported by the UK Health Security Agency, the UK Department of Health and Social Care and the Oxford EPSRC Centre for Doctoral Training in Healthcare Data Science (EP/Y035321/1). The funders had no role in the study design, data analysis, data interpretation, or writing of the report. The views expressed in this article are those of the authors and not necessarily those of the UK Health Security Agency, the UK Department of Health and Social Care, the NIHR or the EPSRC Centre for Doctoral Training in Healthcare Data Science.

Competing interests

WL reports a grant from Gilead Sciences. WL is a member of the London Joint Working Group for Substance Use and Hepatitis C. The remaining authors declare no competing interests.

Use of generative AI

In accordance with the journal's policy on artificial intelligence, we confirm that generative AI tools were used both as a component of the study methodology and in the preparation of this manuscript, and disclose the following.

As part of the study methodology, described in full in the Methods: ARBITER performs risk-of-bias assessment using a large language model (Gemini 2.5 Pro, Google) as its inference backend; Claude Sonnet 4.6 (Anthropic) served as a third independent assessor; and Claude Sonnet 4.6 generated draft consensus statements and the failure-type analysis from the three assessors' outputs. All consensus judgments and failure classifications were reviewed by the author team, with Substantive and Major discrepancies adjudicated by a human expert panel (W.L., J.S., Z.H.) and W.L. holding oversight of final adjudication decisions. The initial statistical analysis was performed by Claude Sonnet 4.6 and independently verified by a co-author with statistical expertise (J.S.).

In the development of the system, Claude Sonnet 3.7–4.5 (Anthropic) was used as a coding and authoring assistant for the pipeline code and criteria files, with comparative edge-case testing via Gemini 2.5 Pro (Google).

In the preparation of this manuscript, Claude Sonnet 4.6 (Anthropic) was used to support structural revision, tightening of section prose for clarity and length, and drafting of specific sections. All AI-assisted text was reviewed, edited, and approved by the author team.

No AI tool was used as a substitute for author judgment. The study design, methodological framework, interpretation of findings, and conclusions remain those of the authors, who take full responsibility for the integrity and accuracy of the work.

References

1. UNAIDS. Global HIV & AIDS statistics — Fact sheet. 2026.
2. (WHO) WHO. Global hepatitis report 2026. 2026.
3. Wolff RF, Moons KGM, Riley RD, Whiting PF, Westwood M, Collins GS, et al. PROBAST: A Tool to Assess the Risk of Bias and Applicability of Prediction Model Studies. *Ann Intern Med*. 2019;170(1):51-8.
4. Moons KGM, Wolff RF, Riley RD, Whiting PF, Westwood M, Collins GS, et al. PROBAST: A Tool to Assess Risk of Bias and Applicability of Prediction Model Studies: Explanation and Elaboration. *Ann Intern Med*. 2019;170(1):W1-w33.

5. Moons KGM, Damen JAA, Kaul T, Hooft L, Andaur Navarro C, Dhiman P, et al. PROBAST+AI: an updated quality, risk of bias, and applicability assessment tool for prediction models using regression or artificial intelligence methods. *Bmj*. 2025;388:e082505.
6. Langenhuijsen LFS, Groenwold RHH, Cannegieter SC, Klok FA, van Diepen M, de Jong Y. Concordance analysis of the short-form of the PROBAST as a screening tool: two-step risk of bias assessment showing only underestimation of risk of bias. *J Clin Epidemiol*. 2026;194:112232.
7. Hasan B, Saadi S, Rajjoub NS, Hegazi M, Al-Kordi M, Fleti F, et al. Integrating large language models in systematic reviews: a framework and case study using ROBINS-I for risk of bias assessment. *BMJ Evid Based Med*. 2024;29(6):394-8.
8. Lai H, Ge L, Sun M, Pan B, Huang J, Hou L, et al. Assessing the Risk of Bias in Randomized Clinical Trials With Large Language Models. *JAMA Netw Open*. 2024;7(5):e2412687.
9. Lai H, Liu J, Bai C, Liu H, Pan B, Luo X, et al. Language models for data extraction and risk of bias assessment in complementary medicine. *NPJ Digit Med*. 2025;8(1):74.
10. Cabello JB, Ruiz Garcia V, Torralba M. Critical appraisal tools for evaluating artificial intelligence in clinical studies: scoping review. *Journal of Medical Internet Research*. 2025;27:e77110.
11. Leber W, Farooq HZ, Panovska-Griffiths J, Larvin H, Baggaley RF, Divall P, et al. Risk prediction models for targeted testing of HIV, hepatitis B and hepatitis C: a systematic review and meta-analysis. *BMC Infect Dis*. 2025;25(1):1648.
12. Cohen J. Weighted kappa: nominal scale agreement with provision for scaled disagreement or partial credit. *Psychol Bull*. 1968;70(4):213-20.
13. Kerschberger B, Aung A, Mpala Q, Ntshalintshali N, Mamba C, Schomaker M, et al. Predicting, Diagnosing, and Treating Acute and Early HIV Infection in a Public Sector Facility in Eswatini. *J Acquir Immune Defic Syndr*. 2021;88(5):506-17.
14. Šuster S, Baldwin T, Verspoor K. Zero- and few-shot prompting of generative large language models provides weak assessment of risk of bias in clinical trials. *Research Synthesis Methods*. 2024;15(6):988-1000.
15. Gandhi S, Shokravi A, Chelliahpillai Y, Balas M. Evaluating large language model performance in Risk of Bias assessments: A cross-sectional validation study. *PLoS One*. 2026;21(7):e0353155.
16. Hevia A, Chintalapati SS, Lai VKW, Nguyen TT, Wong WT, Klassen T, et al. ROBot2: an interactive system and dataset for LLM-assisted clinical trial risk of bias assessment. *arXiv*. 2025.
17. Beber SA, Groff KD, Mange TR, Bram JT, Fabricant PD. Not ready for prime time: limitations of a retrieval-augmented generation large language model in assessing risk of bias in observational studies. *Journal of the Pediatric Orthopaedic Society of North America*. 2025;14:100294.
18. Minozzi S, Cinquini M, Gianola S, Gonzalez-Lorenzo M, Banzi R. The revised Cochrane risk of bias tool for randomized trials (RoB 2) showed low interrater reliability and challenges in its application. *J Clin Epidemiol*. 2020;126:37-44.
19. Minozzi S, Dwan K, Borrelli F, Filippini G. Reliability of the revised Cochrane risk-of-bias tool for randomised trials (RoB2) improved with the use of implementation instruction. *J Clin Epidemiol*.

2022;141:99-105.

20. Wand H, Iversen J, Wilson D, Topp L, Maher L. Developing and validating a scoring tool for identifying people who inject drugs at increased risk of hepatitis C virus infection. *BMJ OPEN*. 2012;2(1).
21. Riley RD, Ensor J, Snell KI, Debray TP, Altman DG, Moons KG, et al. External validation of clinical prediction models using big datasets from e-health records or IPD meta-analysis: opportunities and challenges. *BMJ*. 2016;353:i3140.
22. Steyerberg EW. *Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating*. 2nd ed: Springer; 2019.
23. Steyerberg EW, Vickers AJ, Cook NR, Gerds T, Gonen M, Obuchowski N, et al. Assessing the performance of prediction models: a framework for some traditional and novel measures. *Epidemiology*. 2010;21(1):128-38.
24. Naaktgeboren CA, Bertens LCM, van Smeden M, de Groot JAH, Moons KGM, Reitsma JB. Value of composite reference standards in diagnostic research. *BMJ*. 2013;347:f5605.
25. Efthimiou O, Seo M, Chalkou K, Debray T, Egger M, Salanti G. Developing clinical prediction models: a step-by-step guide. *BMJ*. 2024;386:e078276.
26. Kim MK, Roupheal C, McMichael J, Welch N, Dasarathy S. Challenges in and Opportunities for Electronic Health Record-Based Data Analysis and Interpretation. *Gut and Liver*. 2024.
27. Steele AJ, Denaxas SC, Shah AD, Hemingway H, Luscombe NM. Machine learning models in electronic health records can outperform conventional survival models for predicting patient mortality in coronary artery disease. *PLOS ONE*. 2018;13(8):e0202344.
28. Goldstein BA, Navar AM, Pencina MJ, Ioannidis JPA. Opportunities and challenges in developing risk prediction models with electronic health records data: a systematic review. *Journal of the American Medical Informatics Association*. 2017;24(1):198-208.
29. Gianfrancesco MA, Goldstein ND. A narrative review on the validity of electronic health record-based research in epidemiology. *BMC Medical Research Methodology*. 2021;21:234.
30. National Institute for H, Care E. Hepatitis B and C testing: people at risk of infection. National Institute for Health and Care Excellence; 2012. Contract No.: PH43.
31. Centers for Disease C, Prevention. Screening and Testing for Hepatitis B Virus Infection: CDC Recommendations - United States, 2023. *MMWR Recommendations and Reports*. 2023;72(1).
32. Cartwright EJ, Patel P, Kamili S, Wester C. Updated Operational Guidance for Implementing CDC's Recommendations on Testing for Hepatitis C Virus Infection. *MMWR Morbidity and Mortality Weekly Report*. 2023;72(28).
33. British HIVA, British Association for Sexual H, Hiv, British Infection A. *Adult HIV Testing Guidelines 2020*. British HIV Association; 2020.
34. Branco P, Torgo L, Ribeiro RP. A Survey of Predictive Modelling under Imbalanced Distributions. *arXiv*. 2015:arXiv:1505.01658.

35. de Groot JAH, Bossuyt PMM, Reitsma JB, Rutjes AWS, Dendukuri N, Janssen KJM, et al. Verification problems in diagnostic accuracy studies: consequences and solutions. *BMJ*. 2011;343:d4770.
36. Stahl D. New horizons in prediction modelling using machine learning in older people's healthcare research. *Age and Ageing*. 2024;53:afae201.
37. Balzer LB, Havlir DV, Chamie G, Clark TD, Koss CA, Jain V, et al. Machine Learning to Identify Persons at High-Risk of Human Immunodeficiency Virus Acquisition in Rural Kenya and Uganda. *Clinical Infectious Diseases*. 2020;71(9):2326-33.
38. De Weggheleire A, Buyze J, An S, Thai S, van Griensven J, Francque S, et al. Development of a risk score to guide targeted hepatitis C testing among human immunodeficiency virus patients in Cambodia. *World journal of hepatology*. 2021;13(9):1167-80.
39. Harabor V, Mogos R, Nechita A, Adam AM, Adam G, Melinte-Popescu AS, et al. Machine Learning Approaches for the Prediction of Hepatitis B and C Seropositivity. *Int J Environ Res Public Health*. 2023;20(3).
40. Peebles K, Palanee-Phillips T, Balkus JE, Beesham I, Makkan H, Deese J, et al. Age-Specific Risk Scores Do Not Improve HIV-1 Prediction Among Women in South Africa. *J Acquir Immune Defic Syndr*. 2020;85(2):156-64.
41. Sharma S, Kadiravan T, Banga A. A clinical prediction rule to identify patients with tuberculosis at high risk for HIV co-infection. *INDIAN JOURNAL OF MEDICAL RESEARCH*. 2009;130(1):51-7 WE - Science Citation Index Expanded (SCI-EXPANDED).
42. Tang D, Zhang M, Xu J, Zhang X, Yang F, Li H, et al. Application of Data Mining Technology on Surveillance Report Data of HIV/AIDS High-Risk Group in Urumqi from 2009 to 2015. *COMPLEXITY*. 2018.
43. Wahome E, Fegan G, Okuku HS, Mugo P, Price MA, Mwashigadi G, et al. Evaluation of an empiric risk screening score to identify acute and early HIV-1 infection among MSM in Coastal Kenya. *AIDS*. 2013;27(13).
44. Yin L, Zhao Y, Peratikos MB, Song L, Zhang X, Xin R, et al. Risk Prediction Score for HIV Infection: Development and Internal Validation with Cross-Sectional Data from Men Who Have Sex with Men in China. *AIDS and behavior*. 2018;22(7):2267-76.
45. Fleiss JL. Measuring nominal scale agreement among many raters. *Psychological Bulletin*. 1971(76(5)):378–82.
46. Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics*. 1977;33(1):159-74.
47. Scott H, Liu A, Buchbinder S, Vittinghoff E, Irvin R, Nelson L, et al. Development and Validation of the Personalized Sexual Health Promotion (SexPro) HIV Risk Prediction Model for Men Who Have Sex with Men in the United States. *AIDS and behavior*. 2020;24(1):274-83.

Tables

Tables 1 to 4 are available in the Supplementary Files section.

Figures

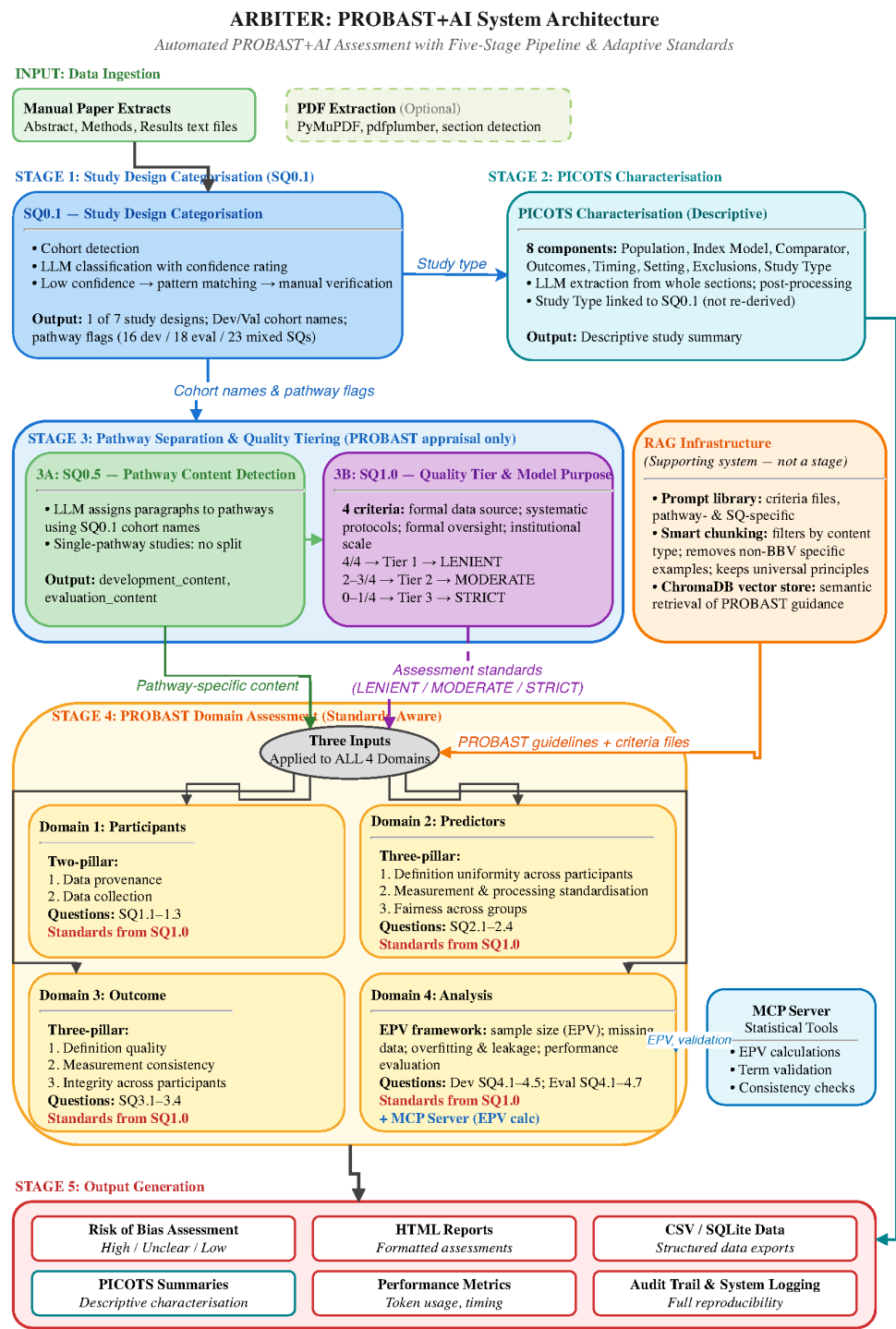


Figure 1

ARBITER five-stage assessment pipeline. In Stage 1 (SQ0.1), the study design is classified, which informs both the descriptive PICOTS characterization in Stage 2 and the PROBAST+AI evaluation pathway. Stage 3 involves the separation of pathway content (SQ0.5) and the assignment of quality tiers that dictate the rigor of the assessment (SQ1.0). Stage 4 evaluates the four PROBAST domains, integrating pathway-specific content, tier-based criteria, and retrieved guidance, with an MCP server aiding in EPV calculations. Stage 5 produces structured outputs. Dashed arrows indicate configuration and context flows, while the RAG infrastructure serves as a supporting system rather than a stage in the pipeline. The Figure has been generated with Claude (Anthropic) and the draw.io (<https://www.drawio.com>) software.

Dynamic Prompt Assembly for PROCAST+AI Signalling Question Assessment

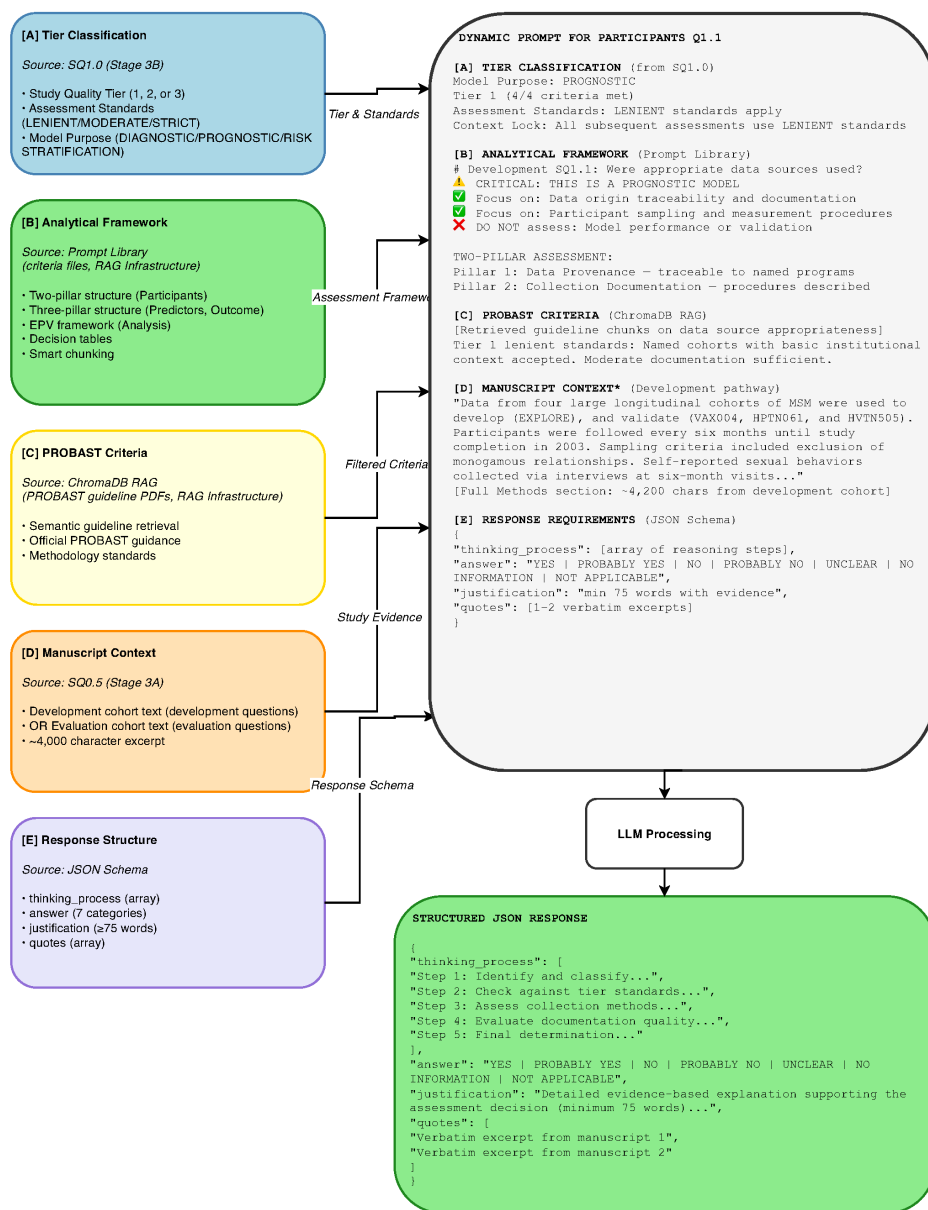


Figure 2

Dynamic prompt assembly for PROCAST+AI signalling question assessment. For each signalling question, ARBITER generates a tailored prompt during execution using five elements [A–E]: tier-based assessment standards (SQ1.0), the analytical framework from the pertinent criteria file, semantically retrieved PROCAST guidance (ChromaDB), manuscript text specific to the pathway (SQ0.5), and the necessary JSON response format. The example provided illustrates the prompt created for Participants

Question 1.1, with the manuscript context excerpt adapted from Scott 2020.(47) Steps 1–5 in the "thinking_process" returned indicate the LLM's internal reasoning for a single question and do not correspond to the five pipeline stages depicted in Figure 1. The Figure has been generated with Claude (Anthropic) and the draw.io (<https://www.drawio.com>) software.

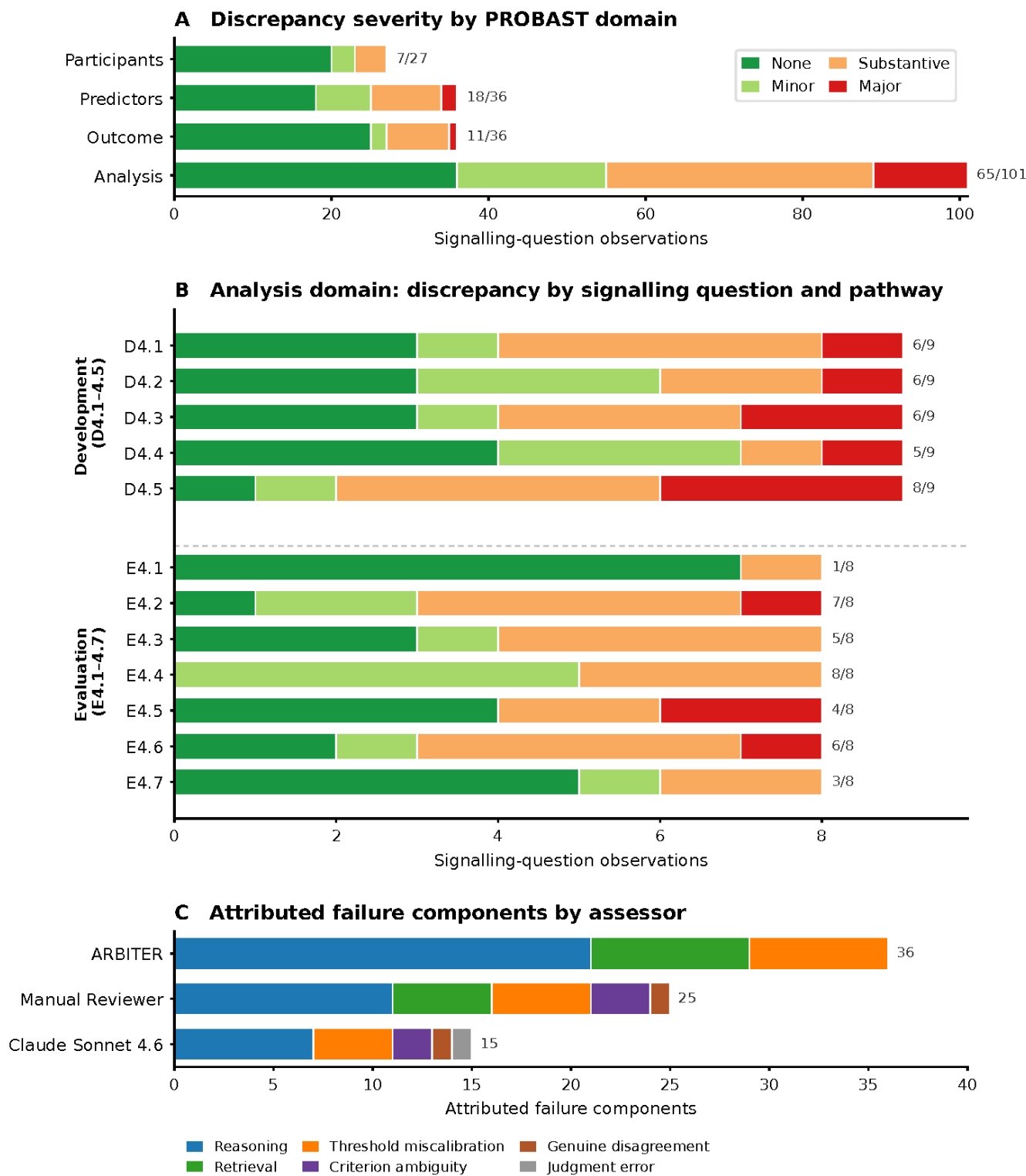


Figure 3

Failure type analysis. Discrepancy and failure patterns across the 200 signalling-question observations of the nine-paper validation set. (A) Discrepancy severity by PROBAST domain and (B) within the Analysis domain by pathway; "None" denotes full three-assessor concordance. (C) Attributed failure components by assessor (ARBITER, Manual Reviewer, Claude Sonnet 4.6); because some discrepancies carry two failure modes across different assessors and unattributed components are omitted, bar totals reflect components rather than the 101 discrepant observations. One Classification cascade failure, identified and resolved during the Kerschberger 2021(13) refinement phase, does not appear. The Figure has been generated with Claude (Anthropic) and the draw.io (<https://www.drawio.com>) software.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [table1to4.docx](#)