Disagreement as a signal: an auditable dual-LLM and human-arbitrated workflow for methodological quality assessment in preclinical complex-intervention evidence

Jianpeng Hou

132515138888@163.com

Shanghai University of Sport

Changhui Wen

Heilongjiang University of Chinese Medicine

Kunpeng Lin

Heilongjiang University of Chinese Medicine

Lingao Xing

Harbin Normal University

Xinyue Yu

Heilongjiang University of Chinese Medicine

Weiyu Liu

Harbin Normal University

Xinyu Li

Heilongjiang University of Chinese Medicine

Yang Li

Heilongjiang University of Chinese Medicine

Research Article

Keywords: Large language models, human arbitration, auditable workflow, evidence synthesis, methodological quality assessment, preclinical animal studies, complex interventions, pre-scoring

Posted Date: July 1st, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10038039/v1

Additional Declarations: No competing interests reported.

Disagreement as a signal: an auditable dual-LLM and human-arbitrated workflow for methodological quality assessment in preclinical complex-intervention evidence

Jianpeng Hou[1,2*†], Changhui Wen[3†], Kunpeng Lin[6], Lingao Xing[4], Weiyu Liu[4], Xinyue Yu[6], Xinyu Li[6], Yang Li[5]

[1*]School of Competitive Sport, Shanghai University of Sport, Shanghai, China.
[2]Department of Physical Education, Heilongjiang University of Chinese Medicine, Harbin, China.
[3]Second Clinical Medical College, Heilongjiang University of Chinese Medicine, Harbin, China.
[4]School of Sports Science, Harbin Normal University, Harbin, China.
[5]School of Humanities, Heilongjiang University of Chinese Medicine, Harbin, China.
[6]Heilongjiang University of Chinese Medicine, Harbin, China.

*Corresponding author(s). E-mail(s): 13251513888@163.com;
†Jianpeng Hou and Changhui Wen contributed equally to this work.

Abstract

Background: Artificial intelligence tools, especially large language models (LLMs), are increasingly entering evidence synthesis workflows. Methodological quality assessment, however, is not merely information extraction. When judgments involve non-specific factors of complex interventions, domain-specific controls, and context-dependent scoring rules, LLM outputs must be embedded within structured provenance, discrepancy detection, and human arbitration. A clearly defined and auditable LLM-assisted workflow for methodological quality assessment in preclinical research remains lacking.

Methods: We developed an auditable dual-LLM pre-scoring workflow. Full-text PDFs were first converted to Markdown using MinerU, and structured evidence

was extracted with source file, original line number, verbatim excerpt, and extraction rationale. Two independent interactive LLM environments then pre-scored the same full texts, evidence tables, and prespecified scoring rules. Discrepant items ($\boldsymbol{\Delta} \geq \mathbf{1}$), ambiguous evidence, and boundary judgments were routed to human arbitration; agreed items were sampled for review to detect shared misclassification; key boundary items underwent additional independent third-party review. We pressure-tested the workflow using 16 animal studies of moxibustion for exercise-induced fatigue and applied adapted scoring versions of SYRCLE (9 items), ARRIVE 2.0 (12 items), and a literature-derived moxibustion-specific framework (MSF; 10 items).

Results: The workflow generated 496 item-level assessments. The two LLM prescores fully agreed on 393 items (79.2%) and disagreed on 103 items (20.8%); all discrepant items underwent human arbitration. Among 50 sampled agreed items, 48 were confirmed and 2 were corrected; this proportion represents sampled confirmation rather than model accuracy. Disagreements concentrated in context-dependent and domain-boundary items, including inclusion/exclusion criteria, outcome reporting, baseline characteristics, specificity interpretation, fixation/environment matching, and sham or non-acupoint controls. Independent third-party review corrected eight systematic scoring inconsistencies. The structured item matrix also revealed recurrent methodological patterns not readily captured by generic tools: insufficient sham or non-acupoint controls, inadequate thermal and fixation/environment matching, incomplete dose reproducibility, and weak safety reporting. In paired construct comparison, MSF percentage scores exceeded SYRCLE scores in all 16 studies (mean $\boldsymbol{\Delta} = +23.9$ percentage points), driven mainly by dose and operational reproducibility.

Conclusions: The methodological contribution of this study is an auditable, human-arbitrated dual-LLM pre-scoring workflow. LLMs were constrained to evidence localization, pre-scoring, and discrepancy exposure rather than autonomous quality assessment. Core components of the workflow–structured provenance, dual-model discrepancy-triggered arbitration, sampled review of agreed items, and independent third-party review–are transferable across domains. The moxibustion pressure test further shows that generic tools and MSF assess different constructs and can be used complementarily to locate methodological weaknesses in complex-intervention research.

Keywords: Large language models; human arbitration; auditable workflow; evidence synthesis; methodological quality assessment; preclinical animal studies; complex interventions; pre-scoring

Corresponding author: Jianpeng Hou, 13251513888@163.com

1 Background

AI tools are rapidly entering medical research methodology, with LLMs increasingly used in systematic review and evidence synthesis workflows. Ruan et al. empirically compared several AI screening tools and showed that false-negative and false-positive profiles differ across tools, so reliability cannot be assumed before deployment [1].

Trad et al. demonstrated how prompt engineering and retrieval-augmented generation can accelerate systematic review screening while retaining human oversight [2]. Mortezaagha et al. extended the problem to full-text biomedical PDFs, emphasizing schema constraints, structured fields, and source traceability in evidence extraction [3]. Taken together, these studies do not simply claim that AI can replace reviewers; rather, they clarify how AI tools may support evidence handling under verifiable sources, prespecified rules, and human supervision.

Methodological quality assessment is more judgment-intensive than literature screening or factual extraction. Lai et al. introduced LLMs into risk-of-bias assessment for clinical randomized trials, providing an important empirical starting point [4]. Forero et al. further compared multiple LLMs on critical appraisal tasks such as ROBIS and AMSTAR 2, showing that model outputs require cautious interpretation even for relatively standardized tools [5]. The difficulty is not merely finding a sentence in the text, but deciding whether an experimental design satisfies a prespecified methodological criterion. For animal experiments and complex interventions, this decision may also depend on domain knowledge, intervention mechanism, control design, and reporting standards. A more realistic question is therefore not whether LLMs can replace human assessors, but whether they can help localize evidence, generate preliminary scores, and expose uncertainty for human arbitration within an auditable evidence chain.

Preclinical animal experiments are an important setting for testing this question. SYRCLE's risk-of-bias tool [6] and the ARRIVE 2.0 reporting guideline [7] are widely used generic instruments for animal research. CRIME-Q further integrates reporting quality, methodological quality, and risk of bias in a unified animal research appraisal tool [8]. These tools are valuable, but their target remains generic appraisal. They can answer whether randomization, blinding, or sample size justification was reported, but they may not answer whether a specific intervention's control condition adequately matches its key non-specific factors.

Moxibustion animal experiments provide a suitable pressure test for this problem. Unlike a single-component pharmacological intervention, moxibustion combines acupoint stimulation, local heat, smoke or odor exposure, and fixation or restraint. If controls do not match these non-specific factors, it becomes difficult to distinguish acupoint-specific effects from thermal effects, smoke effects, and handling stress. Dose parameters, including moxa stick or cone specification, cone count, duration, frequency, treatment course, distance, and temperature, also directly affect reproducibility. Prior methodological evaluations of traditional Chinese medicine animal studies have repeatedly identified weaknesses in blinding, randomization, and sample size justification [9–12], but these generic problems do not fully represent the operational blind spots of moxibustion itself. T/CAAM 0002–2024 provides a reporting guideline for acupuncture and moxibustion animal experiments [13], but it is a reporting checklist rather than an intervention-specific quality assessment tool. In this study, moxibustion is therefore not the primary object of review; it is a complex case used to test whether an LLM-assisted methodological appraisal workflow can handle domain-specific judgments.

On this basis, we aimed to develop and pilot an auditable, human-arbitrated dual-LLM pre-scoring workflow, using moxibustion animal experiments as a pressure-test corpus. We did not attempt to prove that LLMs can autonomously complete quality assessment. Instead, we defined their auxiliary role within a traceable evidence chain. The workflow was applied to 16 animal studies of moxibustion for exercise-induced fatigue to create an auditable item-level scoring matrix. We then evaluated pre-score agreement, discrepancy types, and the value of human arbitration, and used MSF as a case scoring architecture to examine construct differences from SYRCLE.

2 Methods

2.1 Study design and workflow

We designed this work as a methodological pilot study. The core object was an auditable, human-arbitrated dual-LLM pre-scoring workflow; moxibustion animal studies were used as a pressure-test case for complex-intervention quality assessment. The workflow consisted of corpus construction, PDF-to-Markdown conversion, structured evidence extraction with source line numbers, independent pre-scoring by two LLM environments, agreement checking, human arbitration of discrepancies, sampled review of agreed items, third-party independent review, and construction of the final adjudicated scoring matrix. We did not evaluate moxibustion efficacy, and we did not treat LLM output as an automated quality assessment result. The primary focus was how LLMs can support evidence localization, pre-scoring, discrepancy exposure, and cross-study methodological pattern recognition within a traceable evidence chain.

2.2 Corpus source and study screening

We followed PRISMA 2020 for search and screening [14]. The searched databases were PubMed, Embase, The Cochrane Library, Web of Science, CNKI, Wanfang, and VIP. Searches covered database inception to March 2026 and combined concepts related to moxibustion or warm acupuncture with exercise-induced fatigue, exercise fatigue, or exhaustive exercise.

Studies were eligible if they met the following criteria: animal models of exercise-induced fatigue in rats or mice; intervention using pure moxibustion or warm acupuncture in which moxibustion heat stimulation was a core component; at least one control group; and fatigue-related outcomes. Exclusion criteria were non-animal studies, reviews, case reports, conference abstracts, dissertation abstracts, unavailable full text, combined interventions whose independent moxibustion effect could not be separated, non-exercise fatigue models, duplicate publications, or unrecoverable key information. Fig. 1 defines the pressure-test corpus rather than a traditional efficacy review process. The initial search yielded 168 records, 39 full texts were assessed, and the original set of 11 eligible studies was supplemented by an updated search and re-screening that added 5 studies, giving a final corpus of 16 full-text studies and 496 item-level assessments.

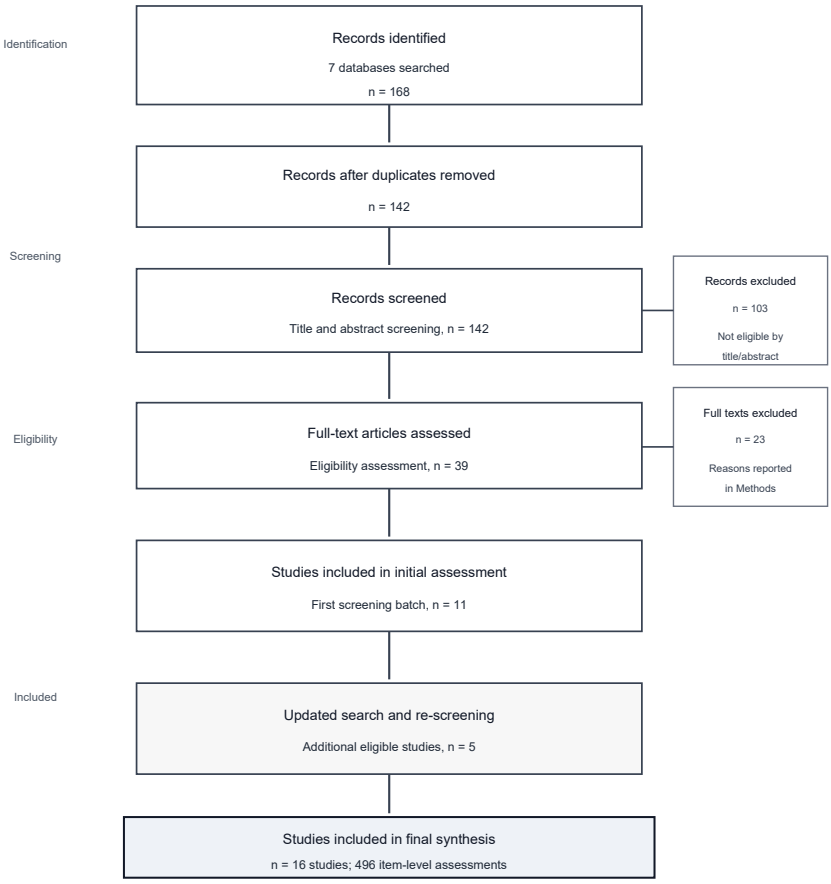

Fig. 1 PRISMA 2020 formation pathway of the pressure-test corpus. The figure shows how 16 full-text studies were formed from the initial search and updated re-screening. All subsequent LLM pre-scoring, human arbitration, and framework comparison used these 16 studies as the analytical corpus.

2.3 PDF conversion, structured evidence extraction, and provenance

Full-text PDFs of included studies were converted to Markdown using MinerU to improve machine readability and reduce the context burden of downstream LLM processing. All subsequent LLM pre-scoring was based on Markdown texts, structured evidence tables, and prespecified scoring rules. When Markdown conversion produced OCR, table, or layout ambiguity, researchers returned to the original PDF for verification.

Structured evidence fields included bibliographic information, animal species and strain, sex, sample size, fatigue model, moxibustion location and dose, intervention frequency and course, control-group design, main outcomes, adverse events, and item-relevant evidence signals for the three appraisal tools. Each evidence record retained

the source file, original line number, verbatim excerpt, and extraction rationale. This design followed the emphasis on schema-constrained and provenance-aware biomedical evidence extraction [3]. In practice, each pre-score was linked back to a structured evidence table and line-numbered source text rather than accepted as a detached model judgment.

For reproducibility, we organized search strategies, included-study information, structured evidence fields, pre-scoring prompts, scoring rules, human arbitration records, third-party review summaries, final scoring matrices, and figure-generation scripts into a reproducibility package. The core analysis chain reconstructs the 496-row item-level audit dataset from raw dual-LLM scoring records and human arbitration records, then generates workflow metrics, locked numerical tables, and figure source data. All reported numbers come from the final adjudicated scoring matrix or reproducible script outputs, not from a single LLM response.

2.4 Dual-LLM pre-scoring, arbitration triggers, and review

Assessment followed a two-step design: LLM pre-scoring followed by human arbitration. For each item, two independent interactive LLM environments processed the same full-text Markdown files, structured evidence tables, and prespecified scoring rules. Each environment produced an initial score, evidence line numbers, a source excerpt, and a scoring rationale. The two pre-scores were then compared. Items with score discrepancies ($\Delta \geq 1$), insufficient evidence, OCR ambiguity, or moxibustion-specific boundary judgments were routed to human arbitration. Items with identical scores and sufficient evidence were treated as provisional and sampled for review by two researchers to check whether agreement concealed shared misclassification. Confidence or review-flag fields were recorded in some batches, but because the fields were not consistently captured across all batches, they were not used as cross-batch quantitative indicators. Fig. 2 emphasizes this arbitration-triggering logic rather than treating LLM outputs as final scores. During arbitration and review, researchers returned to the full-text context, source line numbers, and prespecified scoring rules, correcting misclassification or over-inference under a conservative scoring principle. Unreported or unverifiable information was treated as missing.

2.5 Pressure-test task and case scoring architecture

To provide the dual-LLM workflow with a task that was rule-bound but domain-sensitive, we applied three appraisal architectures: adapted scoring versions of SYRCLE (9 items, 0–18 points), ARRIVE 2.0 core reporting items (12 items, 0–24 points), and a moxibustion-specific framework (MSF; 10 items, 0–20 points). These architectures represented three judgment levels: SYRCLE focused on animal risk-of-bias control, ARRIVE 2.0 on core reporting quality, and MSF on complex-intervention operational quality not directly covered by generic tools. This combination required the same LLM workflow to handle explicit reporting facts, generic methodological judgments, and domain-boundary judgments.

MSF items were derived through an iterative literature-driven process. We first extracted all fragments related to intervention description, control design, and bias

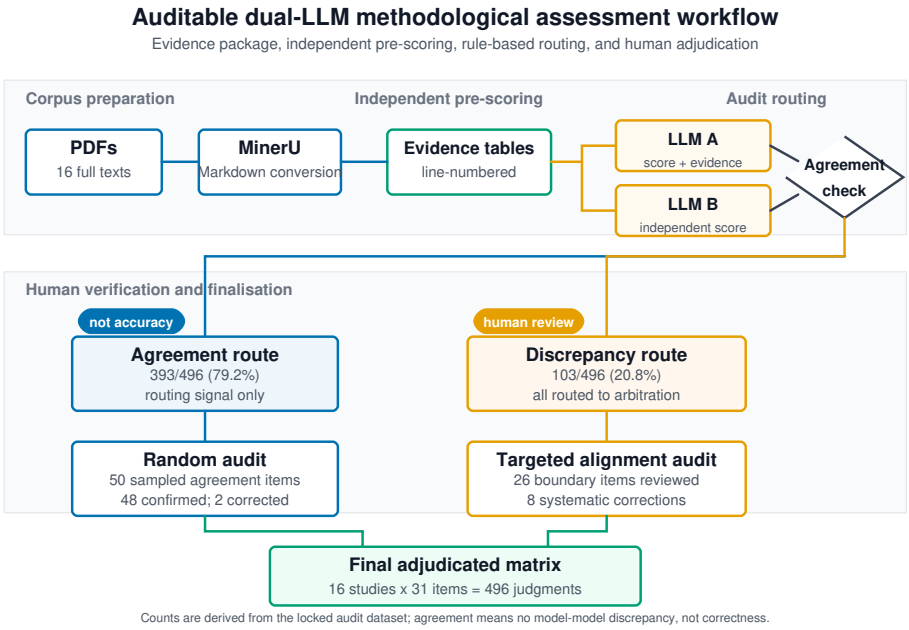

Fig. 2 Auditable dual-LLM methodological assessment workflow. Full-text PDFs were converted to Markdown and structured into line-numbered evidence tables. Two independent LLM environments pre-scored the same evidence package and scoring rules. Agreement checking served only as a routing signal, not an accuracy claim. Discrepant items entered human arbitration, agreed items entered sampled review, key boundary items underwent third-party review, and the final output was a human-confirmed scoring matrix.

control from the 16 included full texts and organized them into 19 candidate items. These items were compared item-by-item with SYRCLE and ARRIVE 2.0. Content heavily overlapping with generic tools, such as sequence generation, data completeness, and statistical methods, was removed. Items that were specific to moxibustion or insufficiently covered by generic tools were retained, and semantically similar items were merged. Two researchers then independently reviewed item definitions, judgment criteria, and scoring rules, resolving disagreements by discussion to form the final 10-item framework.

The final MSF contained three domains and 10 items (Table 1). The control-design domain (T1–T3) assessed whether control groups matched non-specific factors of moxibustion. The dose and operational reproducibility domain (T4–T7) assessed whether intervention parameters were sufficient for replication. The model and safety domain (T8–T10) assessed fatigue-model verification and safety-event reporting. All items were scored from 0 to 2, where 2 indicated adequate and verifiable reporting or design, 1 indicated partial reporting or insufficient information, and 0 indicated absence or clear design inadequacy. In this study, MSF served as a pressure-test scoring architecture and as a source of a comparable item matrix; its reliability, validity, and item weighting require independent external validation.

Table 1 Moxibustion-specific framework (MSF): items, domains, and judgment criteria

Code	Item	Adequate reporting or design criteria
	Domain 1: Control design	
T1	Model control	Normal/blank and model or exercise controls were identifiable, with model procedures and control handling described.
T2	Sham or non-acupoint control	A non-acupoint, tail non-acupoint, sham moxibustion, or equivalent control was specified, with location or procedure described.
T3	Fixation and environment matching	Control and moxibustion groups were clearly matched for fixation time, container, clothing, operator contact, heat, smoke, or odor exposure.
	Domain 2: Dose and operational reproducibility	
T4	Dose reproducibility	Moxa specification, cone count, duration, frequency, course, distance, or temperature were sufficient for replication.
T5	Acupoint localization	Acupoint names and localization basis, anatomical location, or reference standard were provided.
T6	Operator consistency	Same procedure, same device or specification, fixed operator, or standardized handling was described.
T7	Timing consistency	Pre- or post-exercise timing, daily or alternate-day schedule, total course, or total sessions were clear and comparable across groups.
	Domain 3: Model and safety	
T8	Model verification	Exhaustion criteria, treadmill protocol, weight-loading proportion, behavioral indicators, or biochemical validation were reported.
T9	Safety reporting	Presence or absence of adverse events, death, burns, infection, dropouts, or handling of such events was stated.
T10	Specificity interpretation	Results or discussion distinguished or acknowledged moxibustion-specific and non-specific factors.

2.6 Data analysis

Data analysis proceeded at three levels. First, workflow audit metrics described item-level exact agreement, minor discrepancies, and major discrepancies. Exact agreement was defined as identical scores from the two LLMs for the same item; minor discrepancy as a score difference of 1; and major discrepancy as a score difference of 2 or more. Second, discrepancy diagnostics described agreement and disagreement by framework and item, identifying judgments that most often triggered human review. We also summarized the effects of human arbitration, sampled review of agreed items, and third-party review on the final scoring matrix. The sampled review result for agreed items was reported as a confirmation proportion, not as LLM accuracy or gold-standard accuracy.

Third, we analyzed workflow-derived structured insights. For the moxibustion pressure-test case, scores from the three appraisal tools were converted to percentages to remove differences in theoretical maxima: percentage score = actual score / theoretical maximum score $\times 100$. The paired score difference was defined as Δ = MSF percentage score minus SYRCLE percentage score for each study. Because SYRCLE mainly evaluates risk of bias whereas MSF additionally evaluates dose reproducibility,

control adequacy, and safety reporting, this Δ analysis was interpreted descriptively as a construct comparison rather than as evidence that one tool was superior or formally validated. Wilcoxon signed-rank testing was used only as a descriptive paired comparison. Figure and table generation scripts were run from the locked analysis dataset.

3 Results

3.1 Workflow output and pressure-test corpus

The workflow produced 496 item-level assessments across 16 studies and 31 scored items per study (9 SYRCLE, 12 ARRIVE 2.0, and 10 MSF). The two LLM pre-scores agreed on 393 items (79.2%) and disagreed on 103 items (20.8%). Among discrepant items, 98 were minor discrepancies (19.8% of all assessments) and 5 were major discrepancies (1.0%). By framework, exact agreement was 84.7% for SYRCLE (122/144), 78.6% for ARRIVE 2.0 (151/192), and 75.0% for MSF (120/160). The combined generic-tool agreement rate (SYRCLE + ARRIVE 2.0) was 81.2% (273/336).

Table 2 Workflow audit metrics from the final item-level dataset

Metric	Result
Included full-text studies	16 studies
Scored frameworks	SYRCLE (9 items), ARRIVE 2.0 (12 items), MSF (10 items)
Total item-level assessments	496
Dual-LLM exact agreement	393/496 (79.2%)
Any discrepancy	103/496 (20.8%)
Minor discrepancy	98/496 (19.8%)
Major discrepancy	5/496 (1.0%)
Agreed-item sampled review	50 sampled agreed items; 48 confirmed and 2 corrected
Third-party boundary review	26 boundary items reviewed; 8 systematic corrections made

Fig. 3 shows how dual-LLM agreement and disagreement translated into human review work. Agreed items were not treated as true by default; they entered sampled review. Discrepant items were all arbitrated. Sampled review found 2 corrections among 50 agreed items, and third-party boundary review found 8 additional systematic corrections, giving 10 documented corrections to the final scoring matrix.

The final included studies were published from 2006 to 2025. All used male animals; 14 used Sprague-Dawley rats and 2 used Kunming mice. Models included repeated exhaustive swimming, weighted swimming, treadmill fatigue, and long-term exercise training. Common moxibustion sites included CV8, ST36, CV4, and BL23. Intervention forms included mild moxibustion, adhesive moxa cone, seed-sized moxa cone, and lit moxa stimulation. Outcomes covered endurance, biochemical markers, inflammatory factors, oxidative stress, energy metabolism, and tissue morphology.

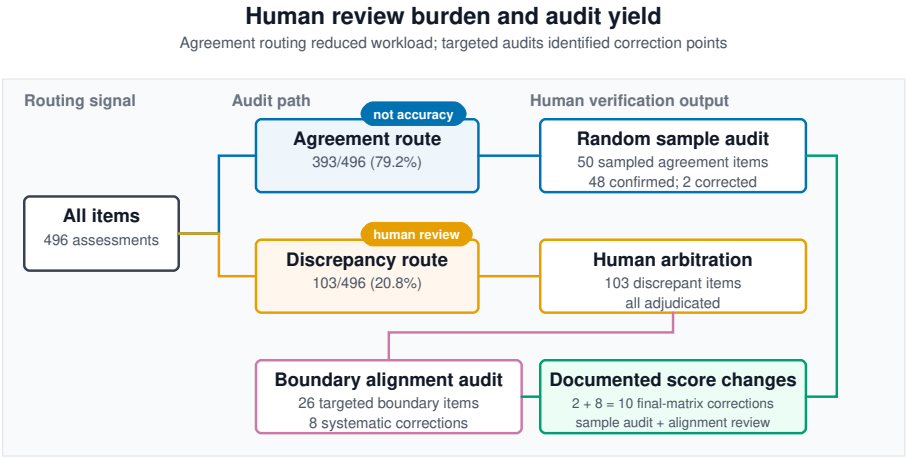

Fig. 3 Human review burden and audit output. Dual-LLM agreement was used as a routing signal for sampled review, not as model accuracy. Dual-LLM discrepancies were all routed to human arbitration. Sampled review identified two corrections, and third-party boundary review identified eight systematic corrections, producing ten documented changes to the final scoring matrix.

Table 3 Overview of included studies

No.	Study	Animal/model	Moxibustion and control overview
1	Liang 2019 [15]	SD rats; repeated exhaustive swimming	Mild moxibustion at CV8 for 15 min, alternate days.
2	Zhang 2025 [16]	SD rats; 8-week treadmill	CV8 moxibustion for 15 min/session, with tail non-acupoint control.
3	Wang 2021 [17]	SD rats; 8-week treadmill	CV8 moxibustion for 15 min/session, with non-acupoint control.
4	Li 2019 [18]	SD rats; 21-day weighted swimming	Adhesive moxa cones at bilateral ST36.
5	Wu 2011 [19]	SD rats; 22-day weighted swimming	Adhesive moxa cones at bilateral ST36.
6	Zhou 2017 [20]	SD rats; repeated exhaustive swimming	Mild moxibustion at CV8 for 15 min, alternate days.
7	Wang 2006 [21]	SD rats; swimming-training fatigue	BL23 moxibustion and non-acupoint control.
8	Gu 2006 [22]	Kunming mice; progressive swimming training	Orthogonal dose design at CV4/ST36/SP6/BL23.
9	Xu 2015 [23]	Kunming mice; 21-day swimming fatigue	Seed-sized moxa cones at ST36 and CV4.
10	Liu 2009 [24]	SD rats; 10-day non-weighted swimming	Bilateral ST36 moxibustion, with acupuncture control.
11	Shen 2021 [25]	SD rats; 12-week treadmill	Mild moxibustion at CV8, with tail non-acupoint control.
12	Li 2019 hippocampus [26]	SD rats; chronic exhaustive swimming	ST36 moxibustion, focused on hippocampal inflammation.
13	Zhou 2017b [27]	SD rats; varying exhaustive swimming	Mild moxibustion at CV8, repeated exhaustion subgroups.
14	Zhang HR 2017 [28]	SD rats; exhaustive exercise	Preventive moxibustion, focused on myocardium.
15	Lu 2011 [29]	SD rats; exhaustive swimming	Moxibustion, focused on liver glycogen and ultrastructure.
16	Mo 2015 [30]	SD rats; exhaustive swimming	Lit moxa stimulation at CV4, focused on skeletal muscle.

3.2 Dual-LLM agreement and discrepancy patterns

Exact agreement was higher for items with explicit evidence or systematically absent evidence. Items such as ARRIVE A1 study design, A2 sample size, A4 randomization, A5 blinding, SYRCLE S1 sequence generation, S3 allocation concealment, S5 performance blinding, MSF T1 model control, and T7 timing consistency reached 100% agreement. Lower agreement occurred for items requiring threshold or contextual interpretation, including ARRIVE A3 inclusion/exclusion criteria (18.8%, 3/16), A6 outcome measures (31.2%, 5/16), A10 result reporting (50.0%, 8/16), SYRCLE S2 baseline characteristics (50.0%, 8/16), MSF T10 specificity interpretation (50.0%, 8/16), T3 fixation/environment matching (62.5%, 10/16), and T2 sham or non-acupoint control and T6 operator consistency (both 68.8%, 11/16).

Fig. 4 converts these stratified results into a diagnostic view. The 10 most frequent disagreement items were A3 (13 disagreements), A6 (11), A10 (8), S2 (8), T10 (8), T4 (7), T3 (6), T2 (5), T6 (5), and S7 (4). These items share a common feature: they require reviewers to decide whether partially reported information is sufficient, whether missing information can be inferred from context, or whether a control condition truly matches domain-specific non-specific factors.

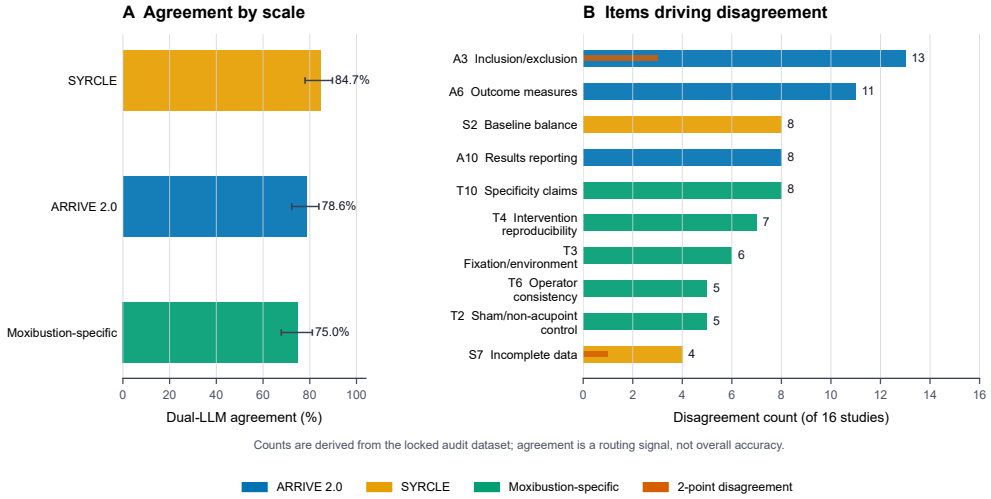

Fig. 4 Dual-LLM agreement and disagreement-concentrated items. Panel A shows exact agreement rates for the three scoring architectures with Wilson 95% confidence intervals. Panel B shows the 10 items with the highest disagreement counts. Disagreement concentrated in context-dependent and domain-boundary judgments, such as inclusion/exclusion criteria, outcome measures, baseline characteristics, specificity interpretation, and moxibustion control design. Agreement rates indicate pre-scoring routing signals, not overall accuracy.

3.3 Human arbitration, sampled review, and third-party review

All 103 discrepant items underwent human arbitration. Across all discrepancies, the final decision matched LLM A in 76 items (73.8%), matched LLM B in 24 items (23.3%), and differed from both LLM pre-scores in 3 items (2.9%). This pattern shows that arbitration was not a simple averaging of model outputs; reviewers returned to the evidence chain and scoring rules for a new judgment. In the updated batch of 5 studies, 27 disputed items were handled by four independent human arbitrators. The final score adopted LLM A's pre-score in 12/27 items (44.4%), LLM B's pre-score in 13/27 items (48.1%), and a score different from both in 2 items.

Agreed items were also not accepted unconditionally. In the sample of 50 agreed items reviewed by two researchers, 48 were retained and 2 were corrected. Therefore, 96.0% should be interpreted as the confirmation proportion within the sampled agreed items, not as the accuracy of the entire dataset. Independent third-party review further identified and corrected eight systematic scoring inconsistencies: six for A3 inclusion/exclusion criteria, one for T2 sham or non-acupoint control, and one for T3 fixation/environment matching.

3.4 Workflow-derived intervention-specific methodological insights

The structured scoring matrix revealed a set of intervention-operational problems that generic tools do not readily capture. For control design, all 16 studies included some form of model or exercise control, but only 4/16 included a non-acupoint control. No study clearly reported a sham moxibustion control or an independent thermal-matching control. Fixation and environment matching were also insufficiently reported; many studies did not state whether control groups received equivalent fixation, container exposure, operator contact, or heat/smoke/odor environment.

For dose and operational reproducibility, most studies reported acupoint names, intervention duration, and treatment course, but temperature and distance parameters were seriously underreported. Some studies reported cone count, time per acupoint, or moxa specification, yet did not fully describe local temperature, distance from moxa to skin, fixation method, or operator consistency. In practical terms, researchers often reported where and for how long moxibustion was applied, but omitted key information needed to reproduce the thermal dose precisely.

For model and safety, most studies described the fatigue model or exhaustion criteria, but safety reporting was weak. ARRIVE A11 and MSF T9 both indicated insufficient reporting of deaths, dropouts, burns, infection, or adverse events. Only a few studies clearly stated death or burn/infection status; most did not state whether adverse events were observed. Fig. 5 combines these findings from two perspectives: which items repeatedly showed deficiencies, and whether those deficiencies were isolated to a few studies or appeared as cross-study patterns.

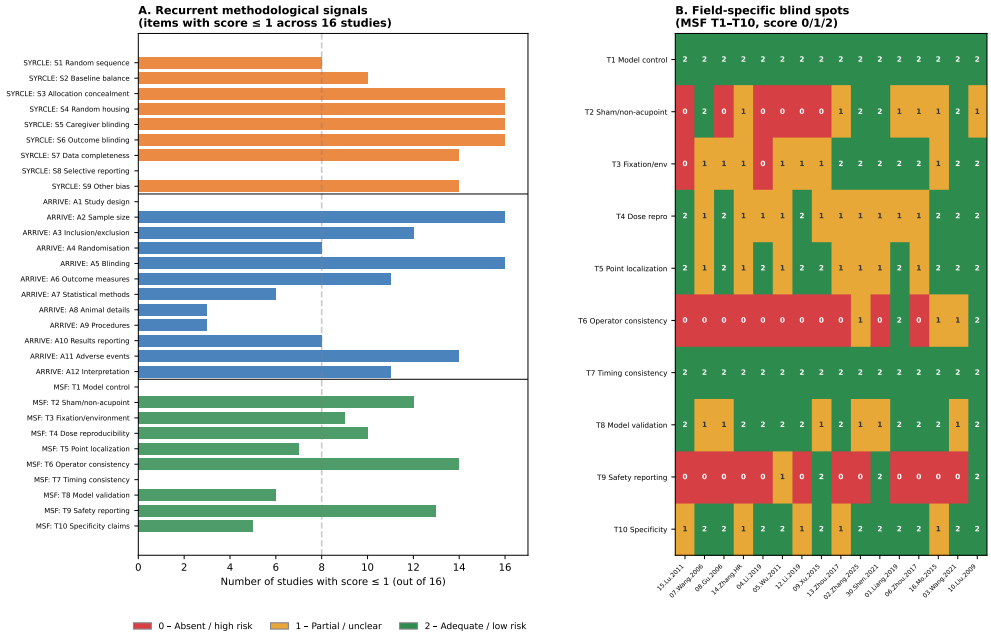

Fig. 5 Recurrent methodological deficiencies and domain-specific control blind spots identified by the workflow. Panel A locates recurrent deficiencies across tools by frequency. Panel B shows the distribution of the 10 MSF items across the 16 studies, indicating that control design, fixation/environment matching, and safety reporting were repeated cross-study patterns rather than isolated cases.

3.5 Construct differences between generic tools and the domain framework

Paired score-difference analysis showed that MSF percentage scores exceeded SYRCLE percentage scores in all 16 studies, with a mean Δ of +23.9 percentage points (Wilcoxon $p < 0.001$); all 16 paired differences were positive, and 14/16 studies had $\Delta > 10$ percentage points. The largest difference occurred in Liu 2009 ($\Delta = +45.0$ percentage points; MSF 95% vs. SYRCLE 50%), and the smallest in Zhou 2017 ($\Delta = +9.4$ percentage points).

Domain decomposition showed that dose and operational reproducibility (T4–T7) was the most stable positive driver, with all 16 studies providing information in this domain that SYRCLE could not evaluate. Control design (T1–T3) showed the greatest between-study variation: a few studies performed relatively well in non-acupoint control or fixation matching, while most scored low. Model and safety (T8–T10) contributed less independently and partly overlapped with generic reporting quality.

The Δ analysis is descriptive. SYRCLE primarily assesses risk of bias, whereas MSF additionally assesses dose reproducibility, control adequacy, and safety reporting. Differences in percentage scores should not be interpreted as evidence of tool superiority or formal validation. The point of Fig. 6 is not which tool scores higher, but where the two assessment constructs differ.

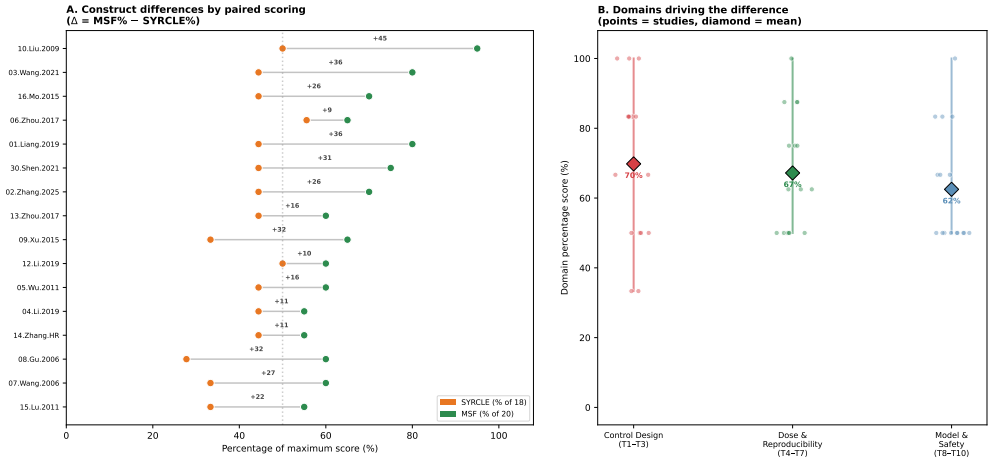

Fig. 6 Construct differences between a generic risk-of-bias tool and MSF. Panel A shows paired percentage-score differences between SYRCLE and MSF for each study. Panel B shows that the difference mainly comes from dose and operational reproducibility, rather than proving that MSF is superior to SYRCLE.

4 Discussion

4.1 Main methodological contribution

This study shows that an auditable dual-LLM workflow can convert full-text evidence from complex-intervention animal studies into a traceable, comparable, and arbitrated item matrix. Its value is not limited to reducing manual reading burden or producing preliminary scores. More importantly, it structures dispersed methodological clues from full texts so that model discrepancies, human review, and cross-study pattern recognition occur within the same evidence chain.

We used LLMs inside a structured evidence chain, not as independent assessors detached from context. This design is continuous with the caution shown by Ruan et al. in externally evaluating AI screening tools [1], the human-supervised automation design of Trad et al. [2], and the emphasis by Mortezaagha et al. on traceable full-text evidence extraction [3]. The present study extends this auditable logic to a more judgment-dependent task: methodological quality assessment.

4.2 Disagreement exposes complex judgment

The dual-LLM pre-scoring workflow achieved 79.2% item-level agreement. This result should be interpreted in relation to existing LLM methodological-appraisal literature. Lai et al. examined LLM judgments for risk of bias in clinical randomized trials [4], and Forero et al. showed that different LLMs do not perform equivalently on critical appraisal of systematic reviews [5]. Compared with those studies, our focus was not to rank a model. Instead, we used the agreement and disagreement between two LLMs

to organize human review. When two models agreed, items could be routed to lower-priority sampled review; when they disagreed, items became high-value audit targets requiring human judgment.

The stratified results show that LLMs more easily agreed on items with explicit evidence or clearly missing evidence, but agreement dropped for items requiring judgments about sufficiency, contextual inference, or matching of intervention-specific non-specific factors. In this sense, LLM disagreement is not merely workflow noise. It is a signal of complex judgment and insufficient evidence. Most disagreements were not simple factual errors; they occurred around partial reporting, OCR fragmentation, boundary information, and moxibustion-specific contextual interpretation. These are exactly the situations where human arbitration is most valuable.

4.3 From pre-scoring to structured methodological insight

The workflow output can go beyond data extraction and item scoring. Once full-text evidence is converted into a line-numbered structured matrix, researchers can compare across studies, items, and appraisal frameworks, identifying repeated methodological patterns that are difficult to capture consistently through single-paper reading alone. The moxibustion pressure test revealed insufficient sham or non-acupoint controls, inadequate fixation/environment matching, incomplete dose reproducibility, and weak safety reporting. These are not autonomous LLM conclusions. They are workflow products: LLMs localize evidence, generate pre-scores, and expose discrepancies; human researchers arbitrate, calibrate scoring thresholds, and interpret cross-study patterns. Thus, the moxibustion findings should not be read as an auxiliary result of a conventional moxibustion review, but as a demonstrative output of an auditable LLM workflow for complex-intervention methodological assessment.

4.4 Transferable workflow and domain-specific rules

Several workflow components are transferable across domains. PDF-to-Markdown conversion and structured evidence extraction can support many full-text evidence tasks when source file, line number, verbatim excerpt, and extraction rationale are retained. Dual-LLM independent pre-scoring is best understood as a discrepancy-detection mechanism rather than a truth-generation mechanism. Routing discrepant items, insufficient evidence, OCR ambiguity, and boundary judgments to human arbitration allows human effort to focus on the parts that most need professional judgment. Sampled review of agreed items and independent third-party review further reduce shared misclassification and scoring-scale drift.

What transfers is the workflow, not the specific items. Complex-intervention quality assessment requires domain customization because key non-specific factors differ by intervention. Moxibustion requires judgments about heat, smoke or odor, fixation/restraint, and acupoint specificity. Acupuncture may require judgments about needling sensation, depth, manipulation, and non-acupoint controls. Surgical models may require judgments about anesthesia, incision, sham surgery, and perioperative management. Reliable application of the workflow in a new domain therefore

requires domain experts to define scoring items, evidence fields, boundary rules, and conservative arbitration principles before LLM pre-scoring begins.

4.5 Implications of the pressure-test case

The moxibustion case shows that generic tools and intervention-specific tools do not assess identical constructs. SYRCLE assesses bias control such as randomization, blinding, and allocation concealment. MSF additionally assesses dose reproducibility, control adequacy, and safety reporting. The mean Δ of +23.9 percentage points indicates that using a generic risk-of-bias tool alone can miss two dimensions central to interpreting moxibustion evidence: whether dose parameters are reproducible and whether controls match heat and fixation/environment factors.

In practical terms, dose and operational reproducibility is the most immediate area for improvement. Temperature, distance, cone count per acupoint, fixation method, and treatment course are usually known during the experiment; what is missing is standardized reporting. Control design is more difficult because sham moxibustion, non-acupoint controls, thermal-matching controls, and fixation matching require coordination among investigators, ethics review, and journal editors. CRIME-Q provides a generic three-dimensional appraisal framework [8], and T/CAAM 0002–2024 provides a reporting checklist for acupuncture and moxibustion animal experiments [13]. The MSF items in this study are best understood as a moxibustion-specific complement to these generic tools and reporting standards.

4.6 Limitations

This study has several limitations. It was piloted in only 16 moxibustion studies of exercise-induced fatigue and cannot prove generalizability to other complex interventions or larger corpora. Because we did not conduct a fully independent dual-human assessment as an expert-adjudicated reference standard, LLM agreement cannot be interpreted as accuracy. For complex-intervention quality assessment, such a reference standard would require at least two trained human reviewers with both methodological and domain expertise to score all items independently, followed by prespecified consensus adjudication and inter-rater reliability assessment. This is resource-intensive because many items involve interpretive judgments about sufficiency, control adequacy, and intervention-specific non-specific factors rather than simple factual extraction. The 96.0% confirmation proportion in the sampled agreed items also cannot replace external reference-standard accuracy. We were unable to quantify time or cost savings relative to a conventional fully manual workflow.

The LLM pre-scoring workflow was implemented through interactive environments rather than pinned API endpoints. Exact model-version control, context-window handling, and deterministic reruns were therefore limited. We mitigated this by using full-text Markdown, structured evidence tables, line-numbered source text, dual-LLM discrepancy checks, and human arbitration, but this does not eliminate the limitation. Some older PDFs also had OCR and line-number continuity problems after PDF-to-Markdown conversion, which may have affected evidence localization.

MSF remains at the pilot stage. Its reliability, validity, and item weighting have not been independently externally validated. Because item generation and pilot application used the same topical corpus, the present results demonstrate preliminary applicability rather than external generalizability. SYRCLE is fundamentally a risk-of-bias judgment tool rather than a summative scale; our 0–18 conversion was used only for descriptive comparison and should not be interpreted as an absolute risk-of-bias grade. No prospective protocol registration was performed, but all analyses followed prespecified scoring rules and extraction templates.

5 Conclusions

We developed and pilot-applied an auditable, human-arbitrated dual-LLM pre-scoring workflow for methodological quality assessment of complex-intervention animal studies. In 16 moxibustion animal studies of exercise-induced fatigue, the two LLMs reached 79.2% exact agreement across 496 items, with disagreements concentrated in context-dependent and domain-specific judgments. The workflow positions LLMs as tools for evidence localization, pre-scoring, discrepancy exposure, and structured comparison, while final judgments remain human-arbitrated. The moxibustion pressure test further shows that structured LLM workflows can help researchers identify recurrent methodological blind spots across studies. Generic risk-of-bias tools and intervention-specific frameworks are complementary; dose reproducibility, control design, and safety reporting are key areas for improvement. Future studies should evaluate reproducibility, generalizability, and practical efficiency in larger samples, additional complex-intervention domains, and with stricter independent human gold standards.

Declarations

Funding

Not applicable.

Availability of data and materials

The reproducibility materials for this study include complete search strategies, included-study information, structured evidence fields, pre-scoring prompts, scoring rules, human arbitration summaries, third-party review summaries, final scoring matrices, workflow audit datasets, locked numerical tables, and figure-generation scripts. The full reproducibility package is publicly available on Zenodo at https://doi.org/10.5281/zenodo.20694413.

Ethics approval and consent to participate

Not applicable. This methodological study was based on published literature and did not involve new animal experiments or human participants.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

J.H. and C.W. jointly conceived and designed the study, developed the MSF, completed literature screening, data extraction, dual-LLM scoring coordination, and drafted the manuscript. K.L. participated in literature screening, data verification, reproducibility-material organization, and manuscript revision. L.X. and W.L. participated in data extraction, item refinement, data checking, and figure and table preparation. X.Y. and X.L. participated in literature screening, data verification, and supplementary-material preparation. Y.L. critically revised the methodological framework, participated in result interpretation, and reviewed the final manuscript. All authors read and approved the final manuscript.

Use of artificial intelligence

The authors declare that generative artificial intelligence tools were used for literature item pre-scoring as described in the Methods section, using two independent interactive LLM environments. All discrepant items underwent human supervision and arbitration, agreed items underwent sampled human review, and key boundary items underwent independent third-party review. Generative AI tools were also used to assist with Python data-analysis and visualization scripts and consistency checking; all script outputs were reviewed by the authors. The authors take full responsibility for the integrity, accuracy, and final content of this manuscript.

References

[1] Ruan M, Fan J, Liu M, Meng Z, Zhang X, Zhang C. Artificial intelligence for the science of evidence synthesis: how good are AI-powered tools for automatic literature screening? BMC Medical Research Methodology. 2025;25(1):199. https://doi.org/10.1186/s12874-025-02644-9.

[2] Trad F, Yammine R, Charafeddine J, Chakhtoura M, Rahme M, El-Hajj Fuleihan G, et al. Streamlining systematic reviews with large language models using prompt engineering and retrieval augmented generation. BMC Medical Research Methodology. 2025;25(1):130. https://doi.org/10.1186/s12874-025-02583-5.

[3] Mortezaagha P, Shaw J, Sun B, Rahgozar A. From chaos to clarity: schema-constrained AI for auditable biomedical evidence extraction from full-text PDFs. BMC Medical Research Methodology. 2026;26(1):119. https://doi.org/10.1186/s12874-026-02847-8.

[4] 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 Network Open. 2024;7(5):e2412687. https://doi.org/10.1001/jamanetworkopen.2024.12687.

[5] Forero DA, Abreu SE, Tovar BE, Oermann MH. Automated analyses of risk of bias and critical appraisal of systematic reviews (ROBIS and AMSTAR 2): a comparison of the performance of 4 large language models. Journal of the American Medical Informatics Association. 2025;32(9):1471–1476. https://doi.org/10.1093/jamia/ocaf117.

[6] Hooijmans CR, Rovers MM, de Vries RBM, Leenaars M, Ritskes-Hoitinga M, Langendam MW. SYRCLE's risk of bias tool for animal studies. BMC Medical Research Methodology. 2014;14:43. Doi: 10.1186/1471-2288-14-43. https://doi.org/10.1186/1471-2288-14-43.

[7] Percie du Sert N, Hurst V, Ahluwalia A, Alam S, Avey MT, Baker M, et al. The ARRIVE guidelines 2.0: Updated guidelines for reporting animal research. PLOS Biology. 2020;18(7):e3000410. Doi: 10.1371/journal.pbio.3000410. https://doi.org/10.1371/journal.pbio.3000410.

[8] Andersen MS, Kofoed MS, Paludan-Müller AS, Pedersen CB, Mathiesen T, Mawrin C, et al. CRIME-Q: a unifying tool for critical appraisal of methodological (technical) quality, quality of reporting and risk of bias in animal research. BMC Medical Research Methodology. 2024;24(1):306. https://doi.org/10.1186/s12874-024-02413-0.

[9] Luo YN, Tian H, Mei ZG, Feng ZT, Yang SB, Fu XY, et al. Acupuncture for experimental cerebral ischemia/reperfusion injury: a systematic review of methodology and reporting quality. Acupuncture in Medicine. 2021;39(6):646–655. https://doi.org/10.1177/09645284211009533.

[10] Jiao L, He X, Zhang J, Liu Y, Luo Y, Wei H. Effects of acupuncture on cancer pain in animal intervention studies: a systematic review and quality assessment. Integrative Cancer Therapies. 2022;21:15347354221123788. https://doi.org/10.1177/15347354221123788.

[11] Li J, Bi X, Hou C, Jin Y, Shang M, Wu X, et al. Methodological quality evaluation of animal experiments on traditional Chinese medicine formulas for glaucoma: A systematic review. European Journal of Integrative Medicine. 2024;71:102399. https://doi.org/10.1016/j.eujim.2024.102399.

[12] Chen R, Zhou X, Deng G, Li S, Li L. Assessment of quality of reporting and methodology in systematic reviews of moxibustion for chronic diseases using PRISMA 2020 and AMSTAR 2. Complementary Therapies in Medicine. 2025;92:103193. https://doi.org/10.1016/j.ctim.2025.103193.

[13] China Association of Acupuncture and Moxibustion. T/CAAM 0002–2024: Guidelines for reporting acupuncture and moxibustion animal experiments. China Association of Acupuncture and Moxibustion; 2024.

[14] Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ. 2021;372:n71. https://doi.org/10.1136/bmj.n71.

[15] Liang YL, Xu XK, Gao F, Wu ZQ, Zhu J, Zhou XH, et al. Effects of moxibustion on central fatigue in rats subjected to different degrees of exhaustive exercise. Physikalische Medizin, Rehabilitationsmedizin, Kurortmedizin. 2019;29:39–44. Doi: 10.1055/s-0044-102011. https://doi.org/10.1055/s-0044-102011.

[16] Zhang ZM, Liu CH, Zuo S, Liu DK, Li RD, Tian YX, et al. The regulatory effect and mechanism of moxibustion at Shenque (CV 8) on spleen immunity in rats with long-term exercise-induced fatigue. World Journal of Traditional Chinese Medicine. 2025;11:225–234. Doi: 10.4103/wjtcm.wjtcm_85_24. https://doi.org/10.4103/wjtcm.wjtcm_85_24.

[17] Wang X. Effect of moxibustion at Shenque on prohibitin 1 expression and energy metabolism in the biceps femoris of exercise-fatigue rats [Master's thesis]. Hebei University of Chinese Medicine; 2021.

[18] Li T. Effects of moxibustion on the IL-6/STAT3 signaling pathway in the frontal cortex of fatigue rats [Master's thesis]. Beijing University of Chinese Medicine; 2019.

[19] Wu Z. Experimental study on the effects of moxibustion on behavior and serum IL-1 beta, IL-6, and TNF-alpha in exercise-fatigue rats [Master's thesis]. Beijing University of Chinese Medicine; 2011.

[20] Zhou LG, Zhou XH, Xu XK, Liang YL, Gao F, Zhang C, et al. Experimental study on the effect of moxibustion at Shenque (CV 8) for long-term exercise-induced fatigue. Journal of Acupuncture and Tuina Science. 2017;15:387–391. Doi: 10.1007/s11726-017-1033-8. https://doi.org/10.1007/s11726-017-1033-8.

[21] Wang ZJ, Wang YH, Liang L, Zhao MH, Wang C, He G, et al. On the effects of moxibustion Shenshu on the resistant ability of rats for exercise-induced fatigue. Journal of Capital Institute of Physical Education. 2006;18(3):52–53.

[22] Gu Y, Jin H, Wu Y, Li S, Ren J. Effect of different moxibustion doses on serum creatine kinase after exercise. Journal of Nanjing University of Traditional Chinese Medicine. 2006;22(6):373–375.

[23] Xu H, Hu Y, Gu Y, Zhang H. Effects of moxibustion with seed-sized moxa cone on apoptosis of myocardial cells after sport fatigue in mice. Chinese Acupuncture and Moxibustion. 2015;35(3):257–263.

[24] Liu HP, Liang B, Zeng CC, Guo ZY. Anti-fatigue effects of acupuncture versus moxibustion at the Zusanli acupoint in exercise-induced fatigue rats. Zhongguo Zuzhi Gongcheng Yanjiu yu Linchuang Kangfu. 2009;13(24):4725–4729. Doi: 10.3969/j.issn.1673-8225.2009.24.026. https://doi.org/10.3969/j.issn.1673-8225.2009.24.026.

[25] Shen ZX, Zhu J, Liang YL, Zhang ZF. Effect of moxibustion on cardiac remodeling and myocardial function in rats with exercise-induced fatigue. World Journal of Traditional Chinese Medicine. 2021;7:254–257. https://doi.org/10.4103/wjtcm.wjtcm_70_20.

[26] Li TG, Shui L, Ge DY, Pu R, Bai SM, Lu J, et al. Moxibustion reduces inflammatory response in the hippocampus of a chronic exercise-induced fatigue rat. Frontiers in Integrative Neuroscience. 2019;13:48. https://doi.org/10.3389/fnint.2019.00048.

[27] Zhou LG, Liang YL, Zhou XH, Zhu J, Gao F, Xu XK, et al. Effect of moxibustion at Shénquè (CV 8) on anti-exercise-induced fatigue in exhaustive rats in varying degrees. World Journal of Acupuncture-Moxibustion. 2017;27(4):35–40. https://doi.org/10.1016/S1003-5257(18)30009-6.

[28] Zhang HR, Zhang HR, Lu SF, Bai H, Gu YH. Effects of preventative moxibustion on AMPK and mTOR in myocardial tissue in rats with exhaustive exercise. Chinese Acupuncture and Moxibustion. 2017;37(5):521–526. https://doi.org/10.13703/j.0255-2930.2017.05.018.

[29] Lu J, Ge GL, Zhang HL, Yin ZZ, Li ZG, Zhang X, et al. Effect of moxibustion on hepatic glycogen and ultrastructure of exercise-induced fatigue rats. Acupuncture Research. 2011;36(1):32–35. https://doi.org/10.13702/j.1000-0607.2011.01.009.

[30] Mo J, Wu LL, Sun ZF, Wang HB, Ding N, He L, et al. Effect of lit-moxa stimulation of Guanyuan (CV 4) acupoint on levels of lactic acid and superoxide dismutase in skeletal muscle of rats. Journal of Traditional Chinese Medicine. 2015;35(2):234–237. https://doi.org/10.1016/S0254-6272(15)30034-0.

Supplementary Files

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

- README.md
- Hou2026dualLLMreproducibilitypackagev1.0.zip
- supplementen.pdf