Supplementary Materials

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*], Lingao Xing[4], Weiyu Liu[4], Xinyue Yu[6], Kunpeng Lin[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

[*] Equal contribution. Corresponding author: 13251513888@163.com

Overview

This supplement provides detailed item-level, study-level, cross-framework, and cross-tool analyses supporting the main manuscript. All analyses are based on dual-LLM initial scoring in two independent interactive environments, followed by human adjudication of divergent items. The main manuscript reports summary statistics; here we present the complete evidence base, including a cross-tool coverage matrix that maps the complementary relationship among existing generic instruments and the moxibustion-specific framework.

Contents:

- **Figure S5:** Evidence extraction quality metrics (16 studies)
- **Figure S6:** Full item-level scoring heatmap (31 items × 16 studies; moved from main text)
- **Figure S7:** Domain decomposition of complementary contribution (16 studies)

Cross-Tool Coverage Matrix

This matrix maps the assessment scope of five instruments relevant to the quality evaluation of moxibustion animal experiments. The purpose is to identify which quality dimensions each tool covers and where the moxibustion-specific framework (MSF) provides **complementary information** not assessed by generic instruments. Symbols: ✓= explicitly covered with dedicated items and scoring criteria; △ = partially or implicitly covered; − = not covered.

Cross-tool coverage matrix. Comparison of five quality assessment instruments across 23 assessment dimensions grouped into three tiers.

Assessment Dimension	SYRCLE	ARRIVE 2.0	CRIME-Q	T/CAAM	MSF
Tier 1: Bias Control (Generic)					
Randomization & sequence generation	✓	✓	✓	✓	−
Baseline comparability	✓	✓	✓	✓	−
Allocation concealment	✓	−	✓	−	−
Random housing	✓	−	−	−	−
Blinding (caregiver & outcome)	✓	✓	✓	−	−
Data completeness	✓	−	✓	−	−
Selective outcome reporting	✓	−	✓	−	−
Tier 2: Reporting Quality (Generic)					
Sample size justification	−	✓	✓	✓	−
Inclusion/exclusion criteria	−	✓	✓	✓	−
Animal details (species, strain, sex, weight)	−	✓	✓	✓	−
Experimental procedures	−	✓	✓	✓	−
Statistical methods	−	✓	✓	✓	−
Results reporting	−	✓	✓	✓	−
Interpretation & limitations	−	✓	✓	✓	−
Adverse events (generic)	−	✓	✓	✓	−
Tier 3: Intervention Quality (Moxibustion-Specific)					
Acupoint specificity control (non-acupoint)	−	−	−	△	✓
Sham moxibustion control	−	−	−	△	✓
Thermal stimulation matching control	−	−	−	−	✓
Smoke/odor environmental control	−	−	−	−	✓
Fixation/restraint matching across groups	−	−	−	△	✓
Dose parameters (cone/specs/duration/frequency)	−	−	−	✓	✓
Temperature and distance parameters	−	−	−	△	✓
Acupoint localization & anatomical basis	−	−	−	✓	✓
Operator consistency & standardization	−	−	−	−	✓
Timing consistency (relative to modeling)	−	−	−	✓	✓
Model validation (fatigue-specific)	−	−	−	✓	✓
Safety reporting (burns, infection, dropout)	−	−	−	△	✓
Specificity rationale (acupoint vs thermal vs smoke)	−	−	−	−	✓

Legend: SYRCLE = SYRCLE Risk of Bias tool for animal studies [1]. ARRIVE 2.0 = Animal Research: Reporting of *In Vivo* Experiments 2.0 [2]. CRIME-Q = Critical appraisal of methodological quality, quality of reporting, and risk of bias in animal research [3]. T/CAAM = T/CAAM 0002–2024 Guidelines for reporting acupuncture and moxibustion

animal experiments [4]. MSF = Moxibustion-specific quality assessment framework (this study).

Key observations: (1) Tier 1 (bias control) and Tier 2 (reporting quality) are comprehensively covered by existing generic instruments but do not address moxibustion-specific operational quality. (2) In Tier 3 (intervention quality), T/CAAM 0002–2024 provides partial coverage via its reporting checklist but lacks dedicated quality-assessment items for thermal matching, smoke/odor control, operator consistency, and specificity rationale. (3) The MSF provides dedicated assessment items with explicit scoring criteria for all 13 Tier-3 dimensions, representing complementary information beyond the scope of generic tools. (4) The MSF intentionally excludes Tier-1 and Tier-2 dimensions to avoid redundancy with SYRCLE and ARRIVE 2.0; it is designed as a complement to, not a replacement for, these instruments.

Supplementary Methods: LLM Scoring Prompt Design

Overview

Each of the 31 items across 16 studies received two independent LLM pre-scores, yielding 496 total item assessments. The pre-scores were generated in two separate interactive environments rather than through pinned API endpoints. Both environments received the same type of inputs: (1) the full-text Markdown of the study, (2) a structured evidence extraction table, and (3) the scoring rubric with item definitions and scoring criteria.

Prompt Structure

The scoring prompt consisted of the following components:

1. **System prompt:** Role assignment as a methodological reviewer specializing in animal experiment quality assessment, with specific expertise in SYRCLE risk of bias, ARRIVE 2.0 reporting guidelines, and TCM external therapy methods.
2. **Framework definitions:** Complete specification of all 31 items across three frameworks, including:
    - Item ID, item name (Chinese + English), and the specific question addressed
    - Scoring criteria: 0 (not reported / high risk / inadequate), 1 (partially reported / unclear), 2 (fully reported / low risk / adequate)
    - Evidence requirements: what constitutes sufficient evidence for each score level
3. **Evidence input:** For each item, pre-extracted evidence lines (line numbers in the Markdown text), direct quotations, and structured extraction table entries.
4. **Output format:** A structured JSON template requiring:
    - `llm_score`: integer 0–2
    - `evidence_lines`: semicolon-separated line numbers referenced
    - `evidence_quote`: verbatim text excerpt supporting the score
    - `rationale`: narrative justification connecting evidence to score
    - `uncertainty`: "low" / "medium" / "high"
    - `needs_human_review`: "yes" / "no"

Adjudication Protocol

After independent scoring:

1. **Agreement check:** Item-level score agreement was calculated across all 496 item assessments (16 studies × 31 items). Perfect agreement (difference = 0) occurred in 79.2% of items (393/496). In the initial batch of 11 studies, agreement was 77.7% (265/341); in the second batch of 5 studies, agreement was 82.6% (128/155).

2. **Minor divergence (difference = 1):** 98/496 items (19.8%) showed a one-point difference between the two LLM pre-scores. These were reviewed against the original text.

3. **Major divergence (difference ≥ 2):** 5/496 items (1.0%; 5/103 discrepant items, 4.9%) showed a two-point difference. These were flagged for additional adjudication and consensus review.

4. **Conservative principle:** Where evidence was ambiguous or OCR quality was poor, the lower score was adopted (or items were marked as "unclear" / 1).

5. **Third-party independent review:** After initial arbitration, a third-party reviewer independently audited all items for systematic scoring inconsistencies across studies, identifying 8 alignment corrections (6 for ARRIVE A3 inclusion/exclusion criteria, 1 for TCM T2 sham/non-acupoint control, and 1 for TCM T3 fixation/environment matching).

Limitations of LLM-Based Scoring

- LLMs may misread context, miss evidence in fragmented OCR text, or over-extrapolate from partial information.
- The LLM pre-scoring workflow used interactive environments rather than pinned API endpoints; exact model-version control, context-window handling, and deterministic reruns were therefore limited.
- The scoring prompt was designed to minimize these risks by requiring explicit line-number citations for every score, but residual errors are possible.
- This workflow should be considered a semi-automated evidence-screening tool rather than a replacement for fully independent dual-human review.
- The high rate of items flagged as "needs_human_review" in the initial batch (204/341, 59.8%) reflects the conservative design of the prompt, which encouraged models to flag borderline cases. In the second batch, a different arbitration workflow was used: 27/155 items (17.4%) were disputed between the two LLMs and routed to four independent human arbitrators, with the final score favoring neither LLM pre-score systematically (48.1% vs. 44.4% of disputes).

Table S1: Item-Level Score Distribution

This table presents the complete distribution of 0/1/2 scores for all 31 items across 16 studies, including the percentage of studies achieving each score level, mean and median scores, and the number of items flagged with high LLM uncertainty or requiring human review.

Table S1: Item-level score distribution (n=16 studies per item). Scores: 0 = not reported / high risk / inadequate; 1 = partial / unclear; 2 = adequate / low risk.

Framework	Item	Label	0 (n)	1 (n)	2 (n)	0%	1%	2%	Mean	High Unc.
SYRCLE	S1	Random sequence	0	8	8	0.0	50.0	50.0	1.50	0
SYRCLE	S2	Baseline balance	3	7	6	18.8	43.8	37.5	1.19	2
SYRCLE	S3	Allocation concealment	16	0	0	100.0	0.0	0.0	0.00	11
SYRCLE	S4	Random housing	1	15	0	6.2	93.8	0.0	0.94	0
SYRCLE	S5	Caregiver blinding	16	0	0	100.0	0.0	0.0	0.00	11
SYRCLE	S6	Outcome blinding	16	0	0	100.0	0.0	0.0	0.00	11
SYRCLE	S7	Data completeness	2	12	2	12.5	75.0	12.5	1.00	3
SYRCLE	S8	Selective reporting	0	0	16	0.0	0.0	100.0	2.00	0
SYRCLE	S9	Other bias	1	13	2	6.2	81.2	12.5	1.06	2
ARRIVE	A1	Study design	0	0	16	0.0	0.0	100.0	2.00	0
ARRIVE	A2	Sample size	0	16	0	0.0	100.0	0.0	1.00	0
ARRIVE	A3	Inclusion/exclusion	0	12	4	0.0	75.0	25.0	1.25	0
ARRIVE	A4	Randomisation	0	8	8	0.0	50.0	50.0	1.50	0
ARRIVE	A5	Blinding	16	0	0	100.0	0.0	0.0	0.00	11
ARRIVE	A6	Outcome measures	0	11	5	0.0	68.8	31.2	1.31	0
ARRIVE	A7	Statistical methods	0	6	10	0.0	37.5	62.5	1.62	0
ARRIVE	A8	Animal details	0	3	13	0.0	18.8	81.2	1.81	1
ARRIVE	A9	Procedures	0	3	13	0.0	18.8	81.2	1.81	2
ARRIVE	A10	Results reporting	0	8	8	0.0	50.0	50.0	1.50	1
ARRIVE	A11	Adverse events	13	1	2	81.2	6.2	12.5	0.31	8
ARRIVE	A12	Interpretation	0	11	5	0.0	68.8	31.2	1.31	0
TCM	T1	Model control	0	0	16	0.0	0.0	100.0	2.00	0
TCM	T2	Sham/non-acupoint	6	6	4	37.5	37.5	25.0	0.88	5
TCM	T3	Fixation/environment	2	7	7	12.5	43.8	43.8	1.31	3
TCM	T4	Dose reproducibility	0	10	6	0.0	62.5	37.5	1.38	2
TCM	T5	Point localization	0	7	9	0.0	43.8	56.2	1.56	0
TCM	T6	Operator consistency	11	3	2	68.8	18.8	12.5	0.44	9
TCM	T7	Timing consistency	0	0	16	0.0	0.0	100.0	2.00	0
TCM	T8	Model validation	0	6	10	0.0	37.5	62.5	1.62	0
TCM	T9	Safety reporting	12	1	3	75.0	6.2	18.8	0.44	9
TCM	T10	Specificity claims	0	5	11	0.0	31.2	68.8	1.69	0

Note: "High Unc." = number of studies where the LLM reported "high" uncertainty for this item. Items with mean score = 0.00 are universally deficient across all 16 studies.

Table S2: Evidence Extraction Quality

Table S2: Per-study evidence extraction metrics. Higher evidence reference counts indicate more text segments were identified and cited during scoring.

Study	Items	Evidence Refs	Avg Refs/Item	With Quote	With Rationale	High Unc.	Needs Review
15 Lu 2011 †	31	45	1.5	31	31	0	2
07 Wang 2006	31	43	1.4	31	31	11	20
08 Gu 2006	31	47	1.5	31	31	10	19
14 Zhang HR †	31	43	1.4	31	31	5	5
04 Li 2019	31	45	1.5	31	31	9	14
05 Wu 2011	31	54	1.7	31	31	9	16
12 Li 2019 †	31	46	1.5	31	31	0	1
09 Xu 2015	31	38	1.2	31	31	8	15
13 Zhou 2017 †	31	44	1.4	31	31	0	0
02 Zhang 2025	31	57	1.8	31	31	7	17
30 Shen 2021	31	43	1.4	31	31	5	17
01 Liang 2019	31	48	1.5	31	31	7	14
06 Zhou 2017	31	48	1.5	31	31	8	15
16 Mo 2015 †	31	43	1.4	31	31	0	1
03 Wang 2021	31	42	1.4	31	31	7	14
10 Liu 2009	31	40	1.3	31	31	5	10
Mean	31	45.4	1.5	31	31	5.7	11.2

Note: All 31 items per study had both evidence quotes and rationales extracted. † = Newly included studies (12–16); discrepant items were adjudicated by four independent human arbitrators. "Evidence Refs" = total number of semicolon-separated line number references across all 31 items. "Avg Refs/Item" = mean references per item. Higher values indicate more granular evidence citation.

Table S3: Study-Level Scores with Domain Breakdown

This table extends the main-text scores by decomposing each framework total into domain-level scores, expressed both as raw scores and as percentages of the domain maximum. Studies are ordered by total quality score (ascending).

Table S3: Study scores decomposed by framework totals and 11 methodological domains. Cells show raw score / % of domain maximum. Domain definitions match Figures S3, S4, and S7.

	Framework Totals			SYRCLE Domains				ARRIVE Domains		
Study	**SYRCLE /18**	**ARRIVE /24**	**MSF /20**	**Rand /4**	**Blind /4**	**Report /4**	**Base /6**	**Design /6**	**OutStat /4**	**Animal /4**
15 Lu 2011 †	6 (33%)	13 (54%)	11 (55%)	1 (25%)	0 (0%)	3 (75%)	2 (33%)	4 (67%)	3 (75%)	3 (75%)
07 Wang 2006	6 (33%)	13 (54%)	12 (60%)	1 (25%)	0 (0%)	2 (50%)	3 (50%)	4 (67%)	2 (50%)	3 (75%)
08 Gu 2006	5 (28%)	14 (58%)	12 (60%)	1 (25%)	0 (0%)	2 (50%)	2 (33%)	4 (67%)	2 (50%)	4 (100%)
14 Zhang HR †	8 (44%)	14 (58%)	11 (55%)	2 (50%)	0 (0%)	3 (75%)	3 (50%)	5 (83%)	3 (75%)	2 (50%)
04 Li 2019	8 (44%)	16 (67%)	11 (55%)	1 (25%)	0 (0%)	3 (75%)	4 (67%)	4 (67%)	3 (75%)	4 (100%)
05 Wu 2011	8 (44%)	15 (62%)	12 (60%)	1 (25%)	0 (0%)	3 (75%)	4 (67%)	4 (67%)	2 (50%)	4 (100%)
12 Li 2019 †	9 (50%)	14 (58%)	12 (60%)	1 (25%)	0 (0%)	3 (75%)	5 (83%)	4 (67%)	3 (75%)	4 (100%)
09 Xu 2015	6 (33%)	17 (71%)	13 (65%)	1 (25%)	0 (0%)	3 (75%)	2 (33%)	4 (67%)	3 (75%)	4 (100%)
13 Zhou 2017 †	8 (44%)	16 (67%)	12 (60%)	2 (50%)	0 (0%)	3 (75%)	3 (50%)	4 (67%)	4 (100%)	3 (75%)
02 Zhang 2025	8 (44%)	16 (67%)	14 (70%)	2 (50%)	0 (0%)	3 (75%)	3 (50%)	5 (83%)	3 (75%)	4 (100%)
30 Shen 2021	8 (44%)	15 (62%)	15 (75%)	2 (50%)	0 (0%)	3 (75%)	3 (50%)	4 (67%)	3 (75%)	3 (75%)
01 Liang 2019	8 (44%)	15 (62%)	16 (80%)	2 (50%)	0 (0%)	3 (75%)	3 (50%)	4 (67%)	3 (75%)	4 (100%)
06 Zhou 2017	10 (56%)	16 (67%)	13 (65%)	2 (50%)	0 (0%)	4 (100%)	4 (67%)	4 (67%)	3 (75%)	4 (100%)
16 Mo 2015 †	8 (44%)	17 (71%)	14 (70%)	2 (50%)	0 (0%)	3 (75%)	3 (50%)	5 (83%)	4 (100%)	4 (100%)
03 Wang 2021	8 (44%)	18 (75%)	16 (80%)	2 (50%)	0 (0%)	3 (75%)	3 (50%)	5 (83%)	3 (75%)	4 (100%)
10 Liu 2009	9 (50%)	18 (75%)	19 (95%)	1 (25%)	0 (0%)	4 (100%)	4 (67%)	4 (67%)	3 (75%)	4 (100%)

	Results /6	**Control /6**	**Dose /8**	**Safety /6**	**Total /62**
Study	**(ARRIVE)**	**MSF Domains**			**All 3**
15 Lu 2011 †	2 (33%)	2 (33%)	6 (75%)	3 (50%)	30 (48%)
07 Wang 2006	3 (50%)	5 (83%)	4 (50%)	3 (50%)	31 (50%)
08 Gu 2006	3 (50%)	3 (50%)	6 (75%)	3 (50%)	31 (50%)
14 Zhang HR †	2 (33%)	4 (67%)	4 (50%)	3 (50%)	33 (53%)
04 Li 2019	4 (67%)	2 (33%)	5 (62%)	4 (67%)	35 (56%)
05 Wu 2011	4 (67%)	3 (50%)	4 (50%)	5 (83%)	35 (56%)
12 Li 2019 †	2 (33%)	3 (50%)	6 (75%)	3 (50%)	35 (56%)
09 Xu 2015	5 (83%)	3 (50%)	5 (62%)	5 (83%)	36 (58%)
13 Zhou 2017 †	3 (50%)	5 (83%)	4 (50%)	3 (50%)	36 (58%)
02 Zhang 2025	2 (33%)	6 (100%)	5 (62%)	3 (50%)	38 (61%)
30 Shen 2021	3 (50%)	6 (100%)	4 (50%)	5 (83%)	38 (61%)
01 Liang 2019	2 (33%)	5 (83%)	7 (88%)	4 (67%)	39 (63%)
06 Zhou 2017	3 (50%)	5 (83%)	4 (50%)	4 (67%)	39 (63%)
16 Mo 2015 †	2 (33%)	4 (67%)	7 (88%)	3 (50%)	39 (63%)
03 Wang 2021	4 (67%)	6 (100%)	7 (88%)	3 (50%)	42 (68%)
10 Liu 2009	6 (100%)	5 (83%)	8 (100%)	6 (100%)	46 (74%)

Domain key: **Rand** = Randomization (S1+S3, max 4); **Blind** = Blinding (S5+S6, max 4); **Report** = Reporting completeness (S7+S8, max 4); **Base** = Baseline/Other (S2+S4+S9, max 6); **Design** = Study design & methods (A1+A2+A3, max 6); **OutStat** = Outcomes & statistics (A6+A7, max 4); **Animal** = Animal & procedures (A8+A9, max 4); **Results** = Results & interpretation (A10+A11+A12, max 6); **Control** = Control design (T1+T2+T3, max 6); **Dose** = Dose & reproducibility (T4+T5+T6+T7, max 8); **Safety** = Model & safety (T8+T9+T10, max 6). † = Newly included studies (12–16). **MSF** = Moxibustion-specific framework. Total maximum = 18 (SYRCLE) + 24 (ARRIVE 2.0) + 20 (MSF) = 62.

Table S4: Domain-Level Aggregate Scores

Table S4: Domain-level mean scores (% of maximum), standard deviation, and range across 16 studies. Domain definitions match Figures S3, S4, and S7.

Framework	Domain (items)	Mean %	SD	Min %	Max %
SYRCLE	Randomisation (S1+S3)	37.5	12.9	25	50
SYRCLE	Blinding (S5+S6)	0.0	0.0	0	0
SYRCLE	Reporting (S7+S8)	75.0	12.9	50	100
SYRCLE	Baseline/Other (S2+S4+S9)	53.1	13.9	33	83
ARRIVE 2.0	Study Design (A1+A2+A3)	70.8	7.5	67	83
ARRIVE 2.0	Outcomes/Stats (A6+A7)	73.4	14.3	50	100
ARRIVE 2.0	Animal/Procedures (A8+A9)	90.6	15.5	50	100
ARRIVE 2.0	Results/Interpret. (A10+A11+A12)	52.1	20.1	33	100
MSF	Control Design (T1+T2+T3)	69.8	22.9	33	100
MSF	Dose/Reproducibility (T4+T5+T6+T7)	67.2	17.0	50	100
MSF	Model/Safety (T8+T9+T10)	62.5	16.7	50	100

Key observation: The SYRCLE Blinding domain shows the lowest mean (0.0%) with zero variance, confirming that blinding is universally absent across all 16 animal moxibustion studies. ARRIVE Animal & Experimental Details shows the highest mean among ARRIVE domains, reflecting consistent reporting of animal characteristics and procedures. The MSF Control Design domain shows the largest variance (SD = 22.9%), reflecting substantial between-study heterogeneity in moxibustion-specific control quality—a dimension not assessed by either SYRCLE or ARRIVE 2.0.

Table S5: LLM Uncertainty Summary

Table S5: Number of studies per item where the LLM reported low, medium, or high uncertainty in its scoring decision (n=16 studies).

Framework	Item	Low	Medium	High
SYRCLE	S1 Random sequence	14	2	0
SYRCLE	S2 Baseline balance	9	5	2
SYRCLE	S3 Allocation concealment	5	0	11
SYRCLE	S4 Random housing	0	16	0
SYRCLE	S5 Caregiver blinding	5	0	11
SYRCLE	S6 Outcome blinding	5	0	11
SYRCLE	S7 Data completeness	2	11	3
SYRCLE	S8 Selective reporting	16	0	0
SYRCLE	S9 Other bias	2	12	2
ARRIVE	A1 Study design	16	0	0
ARRIVE	A2 Sample size	9	7	0
ARRIVE	A3 Inclusion/exclusion	3	13	0
ARRIVE	A4 Randomisation	14	2	0
ARRIVE	A5 Blinding	5	0	11
ARRIVE	A6 Outcome measures	5	11	0
ARRIVE	A7 Statistical methods	10	6	0
ARRIVE	A8 Animal details	13	2	1
ARRIVE	A9 Procedures	12	2	2
ARRIVE	A10 Results reporting	8	7	1
ARRIVE	A11 Adverse events	7	1	8
ARRIVE	A12 Interpretation	6	10	0
MSF	T1 Model control	16	0	0
MSF	T2 Sham/non-acupoint	7	4	5
MSF	T3 Fixation/environment	8	5	3
MSF	T4 Dose reproducibility	13	1	2
MSF	T5 Point localization	10	6	0
MSF	T6 Operator consistency	4	3	9
MSF	T7 Timing consistency	14	2	0
MSF	T8 Model validation	10	6	0
MSF	T9 Safety reporting	7	0	9
MSF	T10 Specificity claims	12	4	0

Note: Uncertainty data are derived from LLM-reported uncertainty levels across the two assessment batches. Items with the highest high-uncertainty counts (S3, S5, S6, A5) correspond to systematically absent evidence that even human reviewers cannot resolve — the LLM is "certain about the absence" rather than uncertain about interpretation. Items with uniformly low uncertainty (S8, A1, T1, T7) correspond to items where evidence is straightforward to locate and classify.

Supplementary Figures

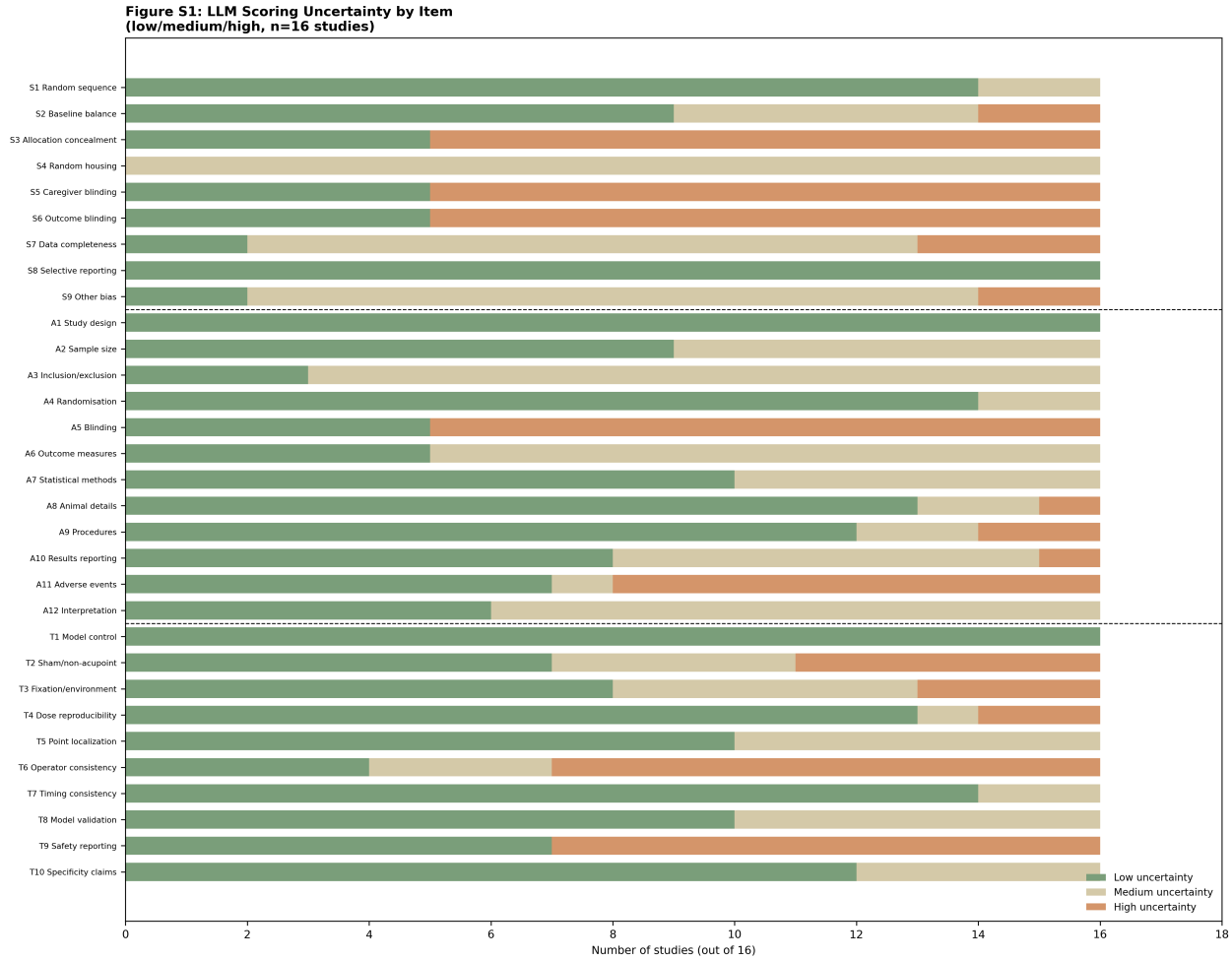

Figure S1: LLM scoring uncertainty distribution by item. Stacked horizontal bars show the number of studies (out of 16) where the LLM reported low (green), medium (gray), or high (amber) uncertainty for each item. Items are grouped by framework (separated by dashed lines). High uncertainty is concentrated in items where evidence is systematically absent (S3 allocation concealment, S5 caregiver blinding, A5 blinding) or where moxibustion-specific judgments require nuanced interpretation (T2 sham/non-acupoint, T6 operator consistency, T9 safety reporting). Uncertainty data combine LLM-reported uncertainty levels across the two assessment batches.

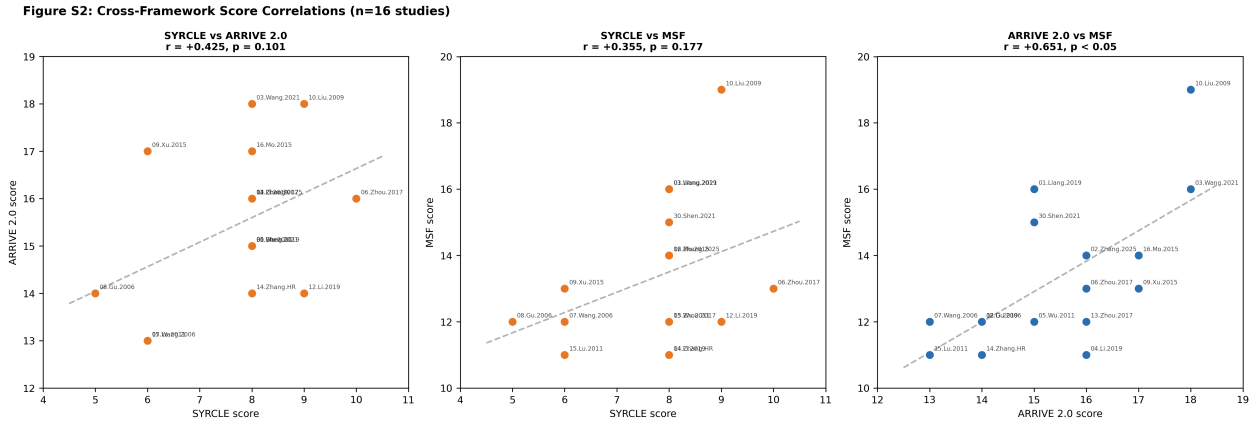

Figure S2: Cross-framework score correlations. Three scatterplots showing the relationship between framework total scores. Each point represents one study (labeled by study number). Dashed lines show linear regression fits. **ARRIVE vs MSF:** r = 0.651, p = 0.006 (significant). **SYRCLE vs ARRIVE:** r = 0.425, p = 0.101 (not significant at $\alpha = 0.05$). **SYRCLE vs MSF:** r = 0.355, p = 0.177 (not significant). The significant ARRIVE–MSF correlation suggests that better-reported studies also tend to have better moxibustion-specific design. The weaker and non-significant SYRCLE–MSF correlation is consistent with the complementary analysis in the main manuscript: the moxibustion-specific framework captures quality dimensions (dose reproducibility, control design, safety) that are structurally distinct from risk-of-bias assessment.

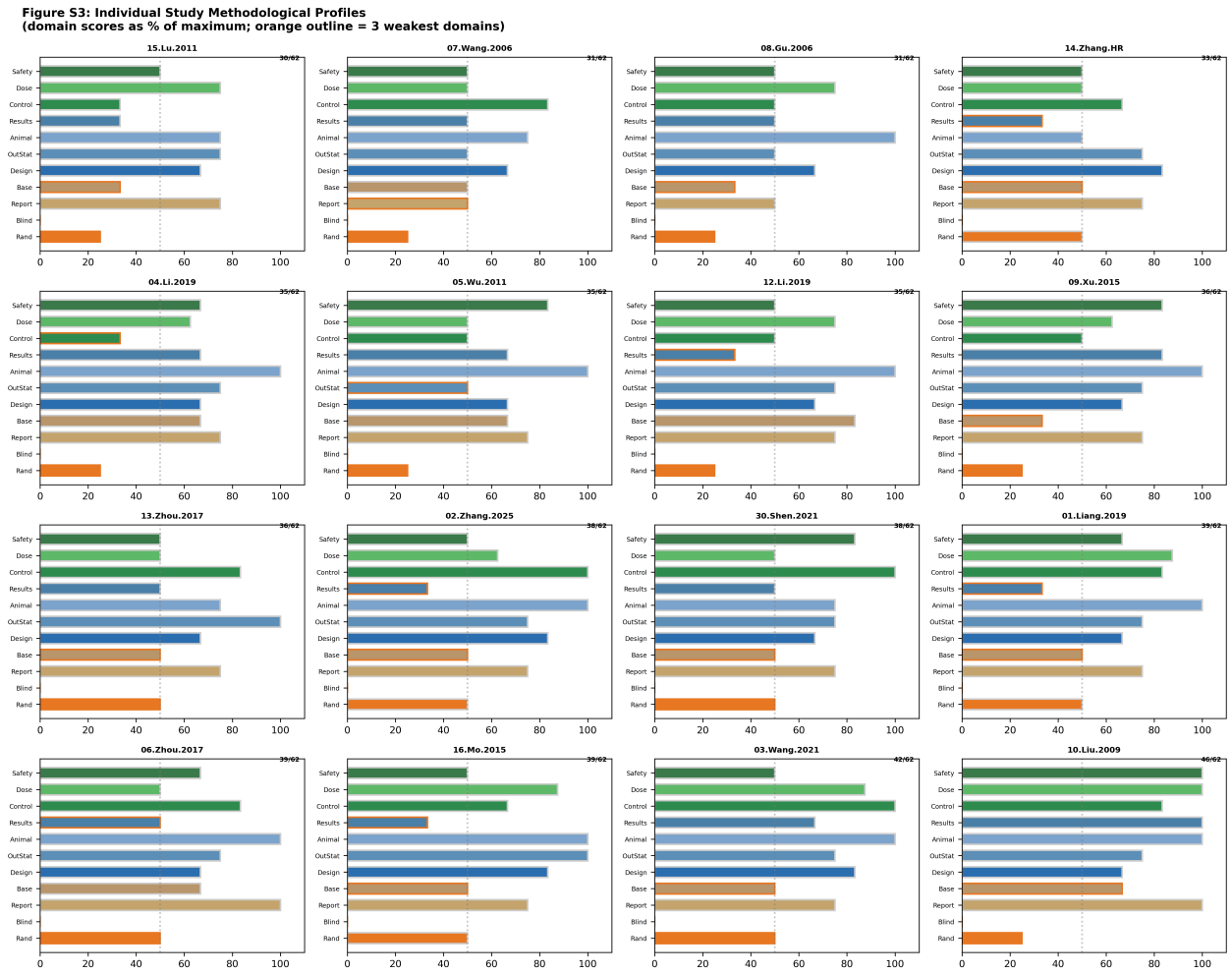

Figure S3: Individual study methodological profiles. Each panel shows one study's domain-level scores as a percentage of the domain maximum. The 11 domains span SYRCLE (Rand = Randomization, Blind = Blinding, Report = Reporting completeness, Base = Baseline/Other), ARRIVE 2.0 (Design = Study design & methods, Out/Stat = Outcomes & statistics, Animal = Animal & experimental details, Results = Results & interpretation), and MSF (Control = Control design, Dose = Dose & reproducibility, Safety = Model & safety). Orange-outlined bars mark each study's three weakest domains. The dashed line at 50% provides a reference threshold. Studies are ordered by total quality score (ascending). MSF = Moxibustion-specific framework.

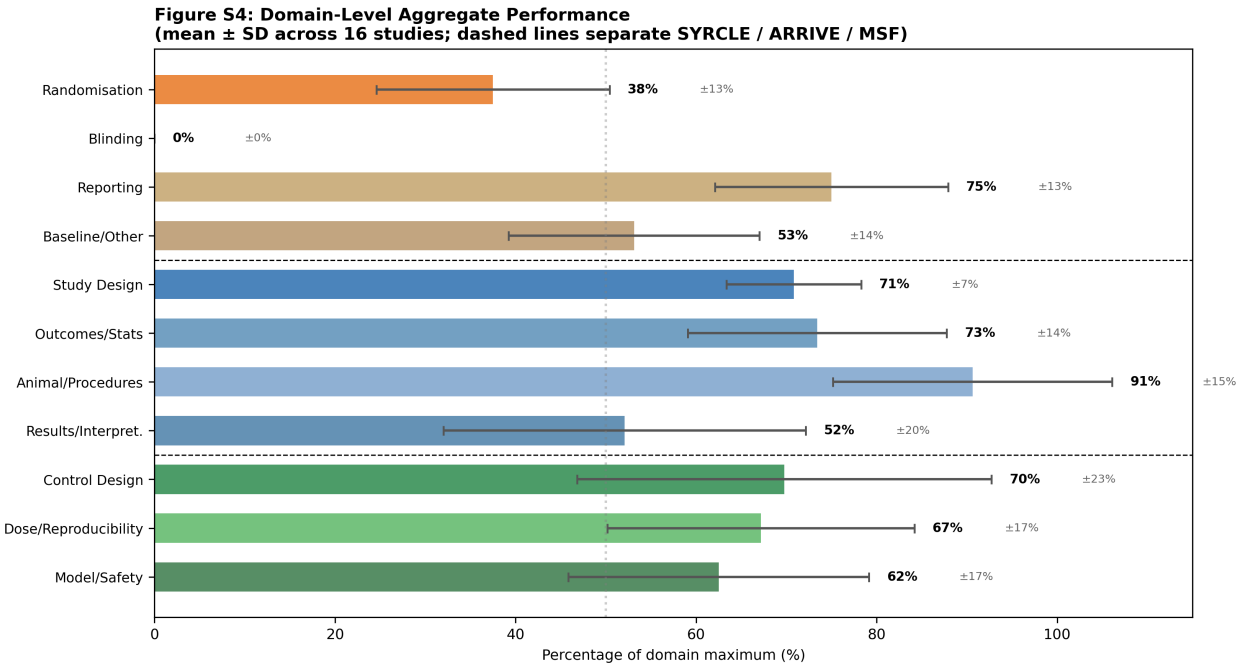

Figure S4: Domain-level aggregate performance. Horizontal bars show the mean percentage of maximum score across 16 studies for each methodological domain, with error bars representing ±1 standard deviation. The Blinding domain (SYRCLE) shows the lowest mean (0.0%) with zero variance, confirming blinding is universally absent across all 16 studies. Animal & Experimental Details (ARRIVE) shows the highest mean (90.6%), reflecting consistent reporting of animal characteristics and procedures. The Control Design domain (MSF) shows the largest variance (SD = 22.9%), indicating substantial between-study heterogeneity in a dimension not assessed by generic tools. MSF = Moxibustion-specific framework.

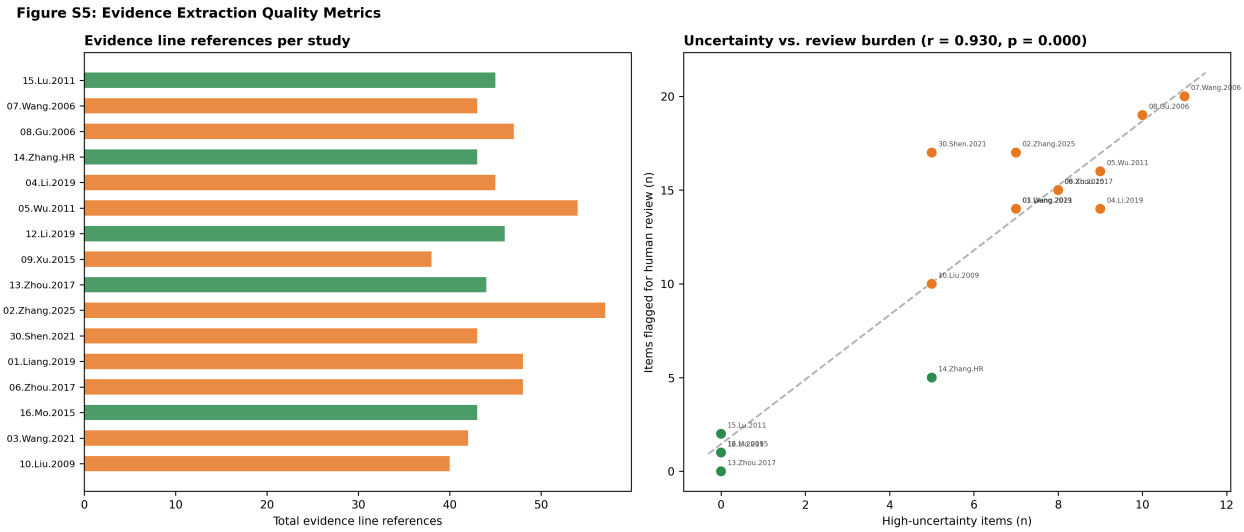

Figure S5: Evidence extraction quality metrics. (Left) Total number of evidence line-number references extracted per study. Studies with more references had more granular text-to-item mapping, which generally corresponds to higher confidence in scoring. **(Right)** Relationship between high-uncertainty item count and items flagged for human review. The strong positive correlation ($r = 0.93$, $p < 0.001$) confirms that LLM uncertainty is robustly associated with borderline cases requiring adjudication, validating the "needs_human_review" flag as a meaningful triage mechanism. Across all 16 studies, study 07 (Wang 2006) shows the highest uncertainty burden (11/31 items), while study 12 (Li 2019-hippo) shows the lowest (0/31 items).

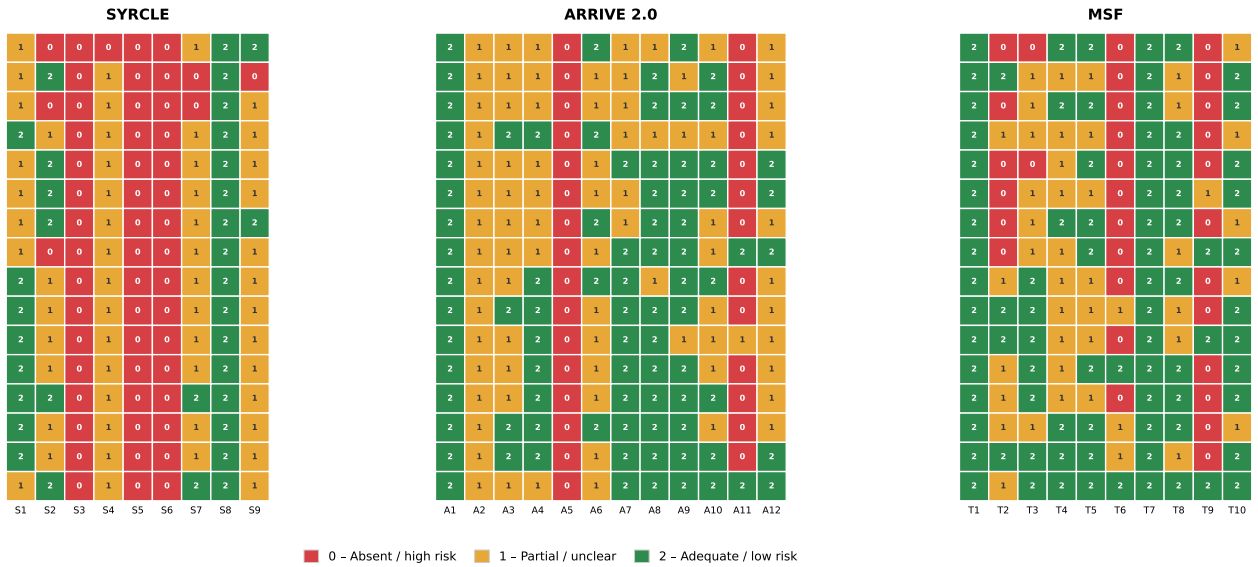

Figure S6: Full item-level scoring heatmap (31 items × 16 studies). Each row is a study, each column an evaluation item. Top section: SYRCLE and ARRIVE 2.0 frameworks; bottom section: moxibustion-specific framework (MSF). Colors represent discrete 0–2 scores (see legend). This figure was originally presented in the main manuscript; it has been moved to the supplement to streamline the main narrative while preserving audit-trail transparency for readers who wish to inspect individual item scores.

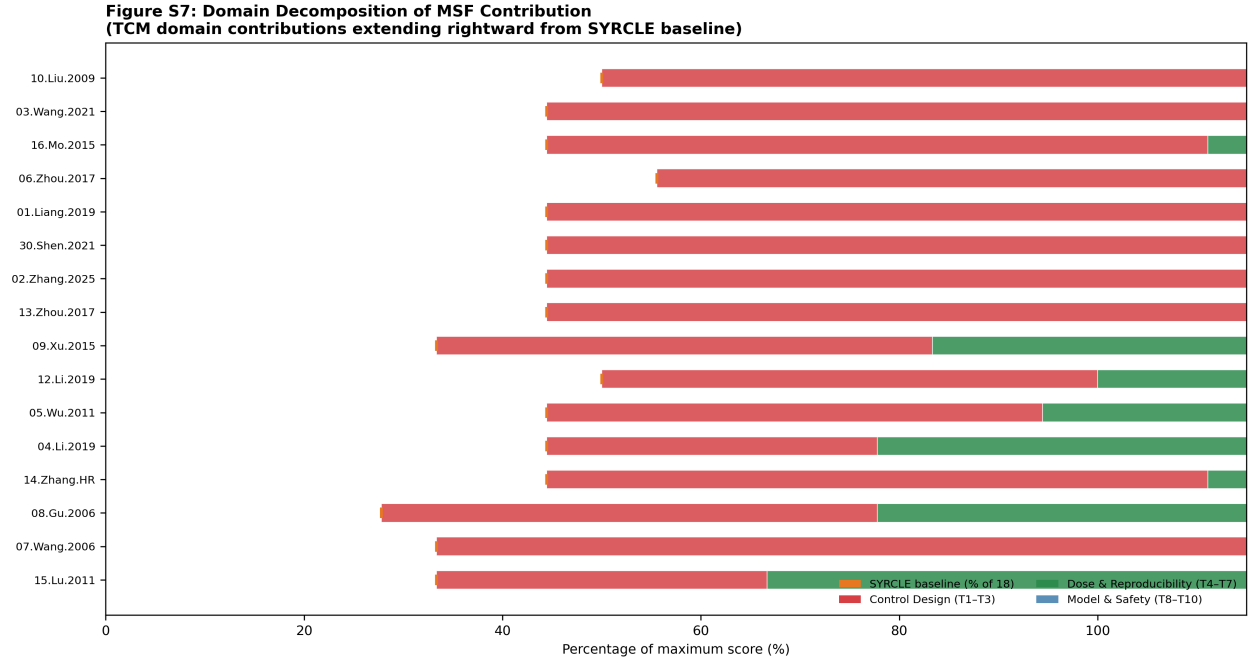

Figure S7: Domain decomposition of the complementary contribution (Figure S7). Stacked bars extend rightward from the SYRCLE baseline (orange vertical tick), showing each of the three moxibustion-specific domains' independent contributions to the score difference (Δ) beyond risk-of-bias assessment. The dose and operational reproducibility domain (T4–T7) is the most consistent driver of the Δ (16/16 studies positive), reflecting the fact that moxibustion dose parameters are naturally recorded by researchers but entirely unassessed by SYRCLE. The control design domain (T1–T3) shows the greatest between-study variability. Score differences are descriptive and reflect differences in assessment content scope; they should not be interpreted as evidence of framework superiority or formal validation.

References

[1] Hooijmans CR, Rovers MM, de Vries RBM, Leenaars M, Ritskes-Hoitinga M, Langendam MW. SYRCLE's risk of bias tool for animal studies. *BMC Med Res Methodol*. 2014;14:43.

[2] Percie du Sert N, Hurst V, Ahluwalia A, et al. The ARRIVE guidelines 2.0: Updated guidelines for reporting animal research. *PLOS Biol*. 2020;18(7):e3000410.

[3] Andersen MS, Kofoed MS, Pedersen AM, et al. CRIME-Q: a unifying tool for critical appraisal of methodological (MQ), quality of reporting (QoR), and risk of bias (RoB) in animal research. *BMC Med Res Methodol*. 2024;24:287.

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