Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Construction of a Large Language Model-Based Literature Evidence Grading System with Dual-Model Adversarial Review and Its Validation in Target Review of Active Components of Lithocarpus litseifolius

Xiaoyong Wang
Nanning Hospital of Traditional Chinese Medicine https://orcid.org/0009-0003-7250-5921
WU Yujuan
The First Affiliated Hospital of Guangxi University of Chinese Medicine https://orcid.org/0009-0001-0132-9700
CHEN Yuanjin
Nanning Hospital of Traditional Chinese Medicine
MAO Weiting
Nanning Hospital of Traditional Chinese Medicine
CHEN Long
Nanning Hospital of Traditional Chinese Medicine
MA Ningning
516252911@qq.com

Nanning Hospital of Traditional Chinese Medicine

Method Article

Keywords: dual-model adversarial review, evidence grading, large language models, Lithocarpus litseifolius, mechanism of action, target library

Posted Date: November 11th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11678943/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: The authors declare no competing interests.

Development and Validation of a Large Language Model–Based Literature Evidence Grading System with Dual-Model Adversarial Review: Application to Active Compounds of *Lithocarpus litseifolius*

WANG Xiaoyong[1], WU Yujuan[2], CHEN Yuanjin[1], MAO Weiting[1], CHEN Long[1], MA Ningning[1]*

(1. Nanning Hospital of Traditional Chinese Medicine (The Affiliated Hospital of Guangxi University of Chinese Medicine), Nanning 530001, China; 2. The First Affiliated Hospital of Guangxi University of Chinese Medicine, Nanning 530001, China)

Abstract

Background

High-quality component–target knowledge is essential for elucidating drug mechanisms, advancing translational research, and supporting precision prescribing. However, existing biomedical databases generally lack systematic evidence grading for component–target associations, while manual review is difficult to scale to the rapidly growing literature.

Objective

We aimed to develop a large language model (LLM)–based literature evidence grading system incorporating dual-model adversarial review and to validate its performance using the active compounds of *Lithocarpus litseifolius* as a case study.

Methods

The system comprised a language model layer and a deterministic rule layer. Two commercial LLMs served as the reviewer and error checker, respectively, forming an adversarial review process; deterministic rules that did not involve language models performed consensus merging and enforced calibration constraints, thereby eliminating stochastic variability at the merging stage. The system incorporated evidence grading criteria (Levels Ia–IV), HUGO Gene Nomenclature Committee (HGNC)–based target standardization, and pathogen-target classification rules. Validation used 776 PubMed records on trilobatin, phloridzin, and phloretin, with manual final review by a researcher as the reference standard. A same-source paired design compared single-model review with dual-model adversarial review.

Results

The error-checking model raised 1,149 challenges, of which 1,120 (97.5%) were adopted; 30 incorrectly listed targets were removed, and 148 and 58 records were downgraded and upgraded, respectively. Compared with single-model review, dual-model review increased

verification accuracy from 78.5% to 84.8% (recall, 95.7%; F1 score, 80.4%) and five-level grading accuracy from 69.8% to 81.6% (Cohen's κ, 0.460 to 0.652; McNemar's test, both $P < 0.001$). The system supported the construction of a high-quality target library comprising 257 traceable records, which were consolidated into 74 unique targets (67 human and 7 pathogen targets); 2, 9, 58, and 5 component–target pairs were classified as Levels Ib, II, III, and IV, respectively.

Conclusion

Dual-model adversarial review combined with deterministic rule calibration significantly improved the accuracy and consistency of literature review and corrected incorrectly listed targets, erroneous mappings, and grading deviations produced by single-model review. This approach provides a reusable framework for scalable, high-quality biomedical knowledge-base construction and evidence-based data governance.

Keywords

dual-model adversarial review; evidence grading; large language models; *Lithocarpus litseifolius*; mechanism of action; target library

Introduction

Component–target interaction knowledge underpins the elucidation of drug mechanisms, translational medicine, and precision prescribing. Its reliability depends on the quality, standardization, and traceability of the underlying evidence [1, 2].

Component–target data in biomedicine are currently derived from two main sources. First, databases such as TCMSP and TCMIP integrate target information through computational prediction or manual curation [3, 4]. Although these resources provide broad coverage, they are updated infrequently, capture the literature incompletely, and, importantly, do not systematically grade the evidence supporting individual targets [5]. Because the strength of evidence for a given target varies substantially across publications, conventional workflows cannot readily distinguish targets supported by systematic reviews or meta-analyses from those supported solely by algorithmic predictions [5]. The indiscriminate inclusion of such targets introduces low-confidence noise into downstream analyses and distorts the resulting target space [6].

Second, manual review by researchers enables quality control but is time-consuming and labor-intensive [7]. Given that a single compound may be associated with hundreds of publications, manual review has become a major bottleneck in the timely updating of knowledge bases [8].

To address these limitations, we developed a large language model–based literature evidence grading system with dual-model adversarial review [9]. The system incorporates evidence grading criteria (Levels Ia–IV), HGNC-based target standardization [10], and pathogen-target classification rules. It supports comparisons of evidence strength, normalization of target names, and complete traceability of the review trajectory for each publication.

We validated the system using the three major active compounds of *Lithocarpus litseifolius*, a medicinal and edible forestry plant native to Guangxi, China: trilobatin, phloridzin, and phloretin [11]. Dual-model adversarial review and target evidence grading were performed to assess the system's performance, with the aim of establishing a reusable approach for the scalable, efficient, and high-quality construction of component–target knowledge bases for traditional Chinese medicine.

1 System Development

1.1 Overall Architecture

The system employs a dual-layer architecture comprising a language model layer and a deterministic rule layer (Figure 1). The language model layer performs the semantic tasks of initial review, error checking, and rebuttal, whereas the deterministic rule layer performs consensus merging, rule-based calibration, standardization, and database entry without invoking a language model, thereby eliminating stochastic variability during merging. The system was implemented in Python, with a Streamlit-based interface for interactive review. Review results are stored in a unified SQLite repository. Each record, indexed by PMID, contains the raw target name, standardized symbol, evidence level, experimental methods, species, confidence score, dual-model review trajectory, and final manual review decision, ensuring end-to-end traceability.

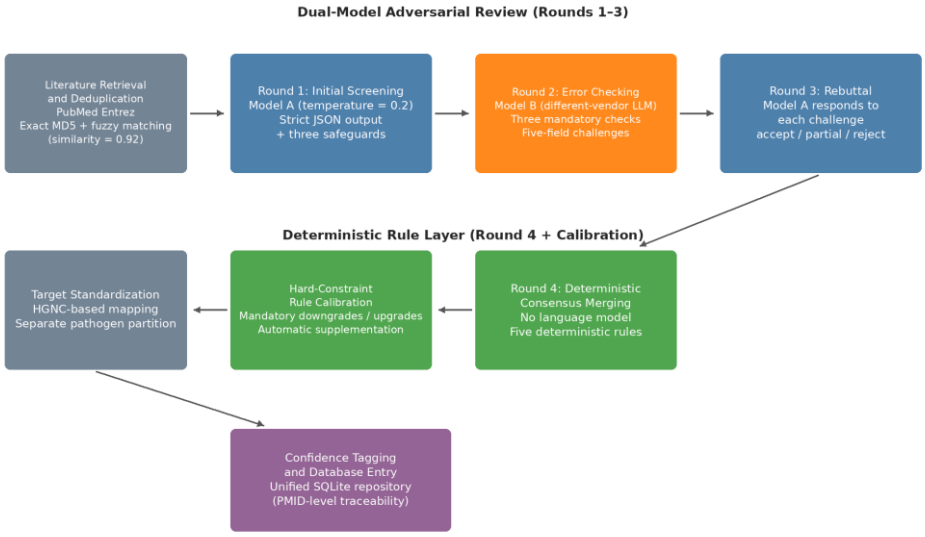


Figure 1. Overall architecture of the literature evidence grading system based on dual-model adversarial review

1.2 Literature Retrieval and Deduplication

PubMed records were retrieved through the NCBI Entrez interface (retmax = 1,000), with titles and abstracts fetched in batches. The search queries included all synonymous spellings of each compound (e.g., trilobatin; phloridzin, phlorizin, and phlorhizin; and phloretin). Publications outside the scope of the study, such as those focusing on plant biosynthesis, were excluded. Search results were cached locally, and repeated queries were served directly from the cache. Deduplication was performed in two steps: exact duplicates were removed using title-based MD5 fingerprints, followed by fuzzy deduplication using a title-similarity threshold of 0.92.

1.3 Round 1: Initial Screening by Model A

Review Model A (a commercial pretrained LLM; temperature = 0.2; max_tokens = 1,500) screened the title and abstract of each publication according to prespecified rules and returned strictly formatted JSON output. The output fields included verification status (is_verified), core direct targets (target_name; up to four per publication), evidence type (evidence_type), evidence level (evidence_level), experimental methods (methods), species, sample size, confidence score (0.25–0.95), and the rationale for the decision. Three safeguards were incorporated into the review rules: (i) species classification—studies reporting in vitro experiments were classified as cellular studies, and in vitro cultures of animal-derived primary cells (e.g., H9c2, PC12, BV2, and primary neurons) were not classified as mammalian in vivo experiments; (ii) constraints on target inflation—downstream pathway effectors (e.g., CAT/SOD/GPx and cleaved caspase-3/9), cell differentiation markers, and inflammatory

cytokines (e.g., IL-6, TNF-α, and IL-1β) could not be listed as core direct targets and were recorded only in the pathway field; and (iii) exclusion of derivatives and compound formulations—studies in which the test substance was a glycosylated derivative, a nanocarrier-loaded formulation, or a compound combination were classified as unverified.

1.4 Round 2: Error Checking by Model B

Error-Checking Model B was a commercial pretrained LLM from a different vendor than Model A, selected to reduce shared blind spots between models [9]. Model B received the title and abstract of each publication together with Model A's complete screening output, including both the original AI-assigned grade and the platform-calibrated grade. It performed three mandatory checks: (i) whether any targets listed by Model A were downstream effector proteins rather than direct targets; (ii) whether any standardized mappings were erroneous, including mappings to pseudogenes or non-coding RNAs, species mismatches, or forced isoform assignments; and (iii) whether the evidence grades assigned by Model A were biased, including underestimation of functional-rescue evidence and assignment of Level II or higher to purely cellular experiments. Each challenge was returned in a structured format comprising issue_type, target_or_field, description, suggested_fix, and severity, enabling precise localization and adjudication.

1.5 Round 3: Rebuttal by Model A

Review Model A responded independently to each challenge with accept, partial, or reject and was required to support its response with text from the abstract. A challenge was adopted when Model A's response was partial and the final_action field contained an explicit instruction such as "remove," "adjust," "correct," or "change to." This rebuttal step prevented unilateral adjudication by the error-checking model. Because challenges raised by Model B could themselves be erroneous, evidence-based rebuttals preserved an auditable adjudication record for each challenge.

1.6 Round 4: Deterministic Consensus Merging and Rule Calibration

Consensus merging did not involve a language model and followed five deterministic rules: (i) when the initial screening and cross-checking results agreed, the initial conclusion was retained; (ii) adopted challenges updated the relevant fields according to suggested_fix; (iii) challenges rejected with supporting textual evidence left the initial conclusion unchanged; (iv) after target removal, the number of targets and the integrity of the evidence chain were reassessed for each publication; and (v) evidence levels were rederived after any field update.

The merged results then passed through a hard-constraint calibration layer comprising mandatory downgrades (10 rules, including that records without mammalian in vivo data could not be graded higher than Level II, animal-derived cell lines could not be graded higher than Level III, and purely enzymatic or biochemical systems could not be graded higher than Level IV), mandatory upgrades (e.g., when functional-rescue evidence had been underestimated), and automatic supplementation (e.g., when protected-set targets had been omitted). Every calibration action was logged.

1.7 Evidence Grading Criteria and Confidence Quantification

The evidence grading criteria and corresponding confidence scores are shown in Table 1. The grading framework is based on evidence-chain strength [12]: direct physical binding combined with functional rescue represents the highest level of evidence, followed by mammalian in vivo experiments and then cellular and in vitro evidence. Molecular docking and molecular dynamics simulations are considered predictive evidence only and cannot support high evidence grades. A literature-level classification of "unverified" excludes the corresponding targets from the database and downstream analyses.

Table 1. Evidence grading criteria and corresponding confidence scores

Level	Criteria	Confidence
Level Ia	Human clinical trials or randomized controlled trials (n ≥ 30)	1.00
Level Ib	Direct physical binding (SPR/MST/CETSA/ITC, etc.), gene overexpression/knockdown functional rescue, and cross-validation by ≥3 methods	0.95
Level II	Mammalian in vivo experiments using ≥2 methods (records without in vivo animal experiments cannot be classified as Level II)	0.85
Level III	Human cell lines with multiple methods, or animal studies with a single method (animal-derived cell lines are capped at Level III)	0.70
Level IV	Human cell lines with a single method, or purely in vitro/enzymatic verification	0.55
Unverified	Literature-level review conclusion; corresponding targets are excluded from the database and downstream analyses	—

1.8 Target Name Standardization

Raw target names generated by the AI models were processed using a standardization module. During preprocessing, parenthetical annotations, Greek letters, and special symbols were removed. Each name was then checked against an in-house alias library and built-in semantic correction rules (e.g., GK→GCK, IRS→IRS1, and XOD→XDH). A family-name safeguard retained generic names such as AMPK and PEPCK when isoform-level evidence was unavailable. Names not resolved by these rules were queried against the mygene.info database (https://mygene.info) and converted to standardized HGNC gene symbols [10]. Proteins encoded by viruses, bacteria, and other pathogens were classified as pathogen targets under the pathogen-target classification rules. Their original names were retained without forced mapping to human genes, their evidence level was capped at Level III, and they were stored in a separate repository partition.

2 Validation Design

2.1 Validation Set

PubMed records on the three major active compounds of *L. litseifolius* constituted the validation set. All three batches were retrieved as described in Section 1.2 and underwent the four-round dual-model review process. After exact and fuzzy deduplication, 52 trilobatin publications entered the review queue. Because their topics were relatively homogeneous, focusing primarily on diabetes and its complications, hepatoprotection, and gut microbiota regulation, this batch served as a small-sample precision validation set. The two larger batches, phloridzin (325 publications) and phloretin (399 publications), covered broader topics and served as large-sample stability validation sets. The three batches comprised 776 records in total, representing 706 unique publications after cross-batch deduplication.

2.2 Reference Standard, Control Setting, and Evaluation Metrics

Manual final review performed publication by publication by a researcher within the platform served as the reference standard. Before making each final judgment, the reviewer could examine the complete dual-model review trajectory. A same-source paired design was used for the control comparison [13]: the Round 1 screening output constituted the Model A single-review arm (same model, prompt, and publications), which was paired with the Round 4 dual-model conclusions generated within the same pipeline, thereby eliminating run-to-run variability. Evaluation metrics included (i) interarm agreement, including agreement rates and Cohen's κ for verification decisions and five-level grading, as well as the direction of decision

changes [14]; (ii) the contribution of adversarial review, including the challenge rate, challenge adoption rate, grading-change rate and direction, and number of incorrectly listed targets removed; (iii) calibration actions, including the numbers of mandatory downgrades, mandatory upgrades, and automatic supplementation events; and (iv) agreement with the reference standard, including verification accuracy, recall, and F1 score [15], five-level grading accuracy and Cohen's κ, and McNemar's paired test for differences between the two arms [16].

2.3 Statistical Analysis

Statistical analyses were performed using Python 3.12 with the scikit-learn [17] and statsmodels [18] packages. Agreement rates were summarized as percentages, and agreement strength was quantified using Cohen's κ [14]. Differences between each review arm and the reference standard were assessed using McNemar's paired chi-square test [16]. Two-sided $P < 0.05$ was considered statistically significant.

3 Results

3.1 Overall Review Results of the Trilobatin Batch

All 52 trilobatin publications completed dual-model review (100% coverage); operational statistics are summarized in Figure 2. Of these publications, 37 (71.2%) were classified as verified, 14 (26.9%) as unverified, and 1 (1.9%) was flagged with a warning. Among the verified publications, 1 was graded Level Ia, 7 Level Ib, and 6 Level II; Levels III and IV together accounted for 23 publications. The single Level Ia record (NFE2L2; PMID 41931990) was revised to Level Ib during manual final review.

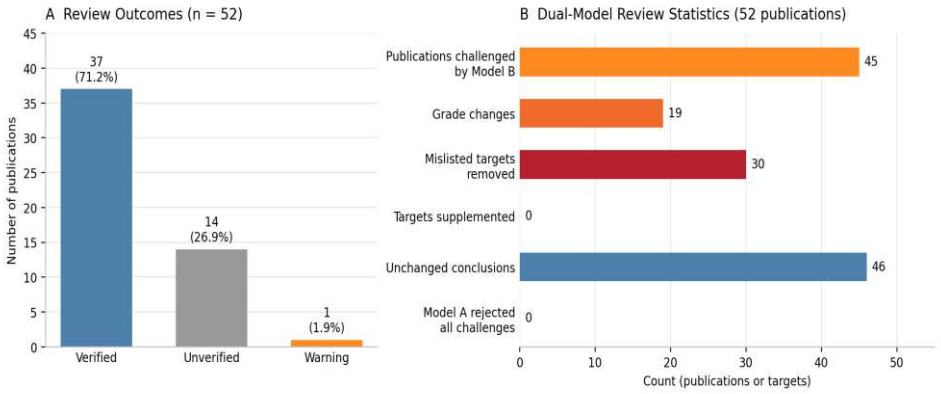


Figure 2. Operational statistics for dual-model adversarial review of the trilobatin batch (n = 52). (A) Distribution of review outcomes. (B) Adversarial review statistics

3.2 Contribution of Adversarial Review

Error-Checking Model B raised 105 structured challenges across 45 publications (86.5%). After rebuttal and deterministic consensus merging, 98 challenges (93.3%) were adopted, resulting in grading changes in 19 publications (36.5%) and removal of 30 incorrectly listed targets. The final conclusions for 46 publications (88.5%) were identical to the initial screening results. No publication had all challenges rejected by Model A. The calibration layer performed 14 mandatory downgrades, 1 mandatory upgrade, and no automatic supplementation. Across all three batches, Model B raised 1,149 challenges, of which 1,120 (97.5%) were adopted. Grading changes comprised 148 downgrades and 58 upgrades (approximately 2.6-fold). One phloridzin publication for which the error-checking interface call failed automatically reverted to the single-model result and was flagged separately; it was excluded from the dual-model statistics.

3.3 Comparison of Model A Single Review and Dual-Model Adversarial Review

The same-source paired comparison of the two review arms is presented in Table 2. Verification decisions agreed in 86.6% of records; the remaining 13.4% (n = 104) changed after dual-model review, with 63 changes to unverified and 41 to verified. Five-level grading agreed in 73.5% of records (Cohen's $\kappa = 0.561$).

Table 2. Same-source paired comparison of Model A single review and dual-model adversarial review

Metric	Trilobatin (n = 52)	Phloridzin (n = 325)	Phloretin (n = 399)	Total (n = 776 records)
Agreement of verification decisions	86.5%	86.5%	86.7%	86.6%
Decision changes (to verified/to unverified)	2/5	17/27	22/31	41/63
Agreement of five-level grading	63.5%	68.0%	79.2%	73.5%
Cohen's κ for five-level grading	0.486	0.490	0.625	0.561
Grading changes (downgrades/upgrades)	7/12	83/21	58/25	148/58

3.4 Comparison with the Manual Final Review Reference Standard

Table 3 compares each review arm with the manual final review reference standard. Across the three batches, dual-model review achieved a verification accuracy of 84.8%, recall of 95.7%, F1 score of 80.4%, five-level grading accuracy of 81.6%, and Cohen's κ of 0.652, compared with 78.5%, 90.1%, 73.2%, 69.8%, and 0.460, respectively, for Model A single review. McNemar's paired tests showed statistically significant differences in verification

decisions (χ^2 = 22.37, P < 0.001) and five-level grading (χ^2 = 51.59, P < 0.001). In the phloridzin and phloretin batches, verification accuracy increased by 7.4 and 6.8 percentage points, respectively, and grading Cohen's κ increased from 0.368 and 0.505 to 0.610 and 0.697, respectively (McNemar's test, both P < 0.001). In the trilobatin batch, the two arms yielded similar results (verification accuracy, 86.5% vs 82.7%; P = 0.688), with differences limited to 2–3 publications.

Table 3. Agreement of the two review arms with the manual final review reference standard

Batch	Review mode	Verification accuracy	Verification recall	Verification F1	Grading accuracy	Grading Cohen's κ
Trilobatin (n = 52)	Model A single review	86.5%	92.3%	91.1%	69.2%	0.536
Trilobatin (n = 52)	Dual-model adversarial review	82.7%	87.2%	88.3%	67.3%	0.533
Phloridzin (n = 325)	Model A single review	75.7%	89.9%	69.3%	63.7%	0.368
Phloridzin (n = 325)	Dual-model adversarial review	83.1%*	97.0%	77.7%	79.4%*	0.610*
Phloretin (n = 399)	Model A single review	79.7%	89.6%	71.8%	74.9%	0.505
Phloretin (n = 399)	Dual-model adversarial review	86.5%*	97.4%	80.6%	85.2%*	0.697*
Total (n = 776 records)	Model A single review	78.5%	90.1%	73.2%	69.8%	0.460
Total (n = 776 records)	Dual-model adversarial review	84.8%*	95.7%	80.4%	81.6%*	0.652*

*Note: Within each batch, dual-model adversarial review was compared with Model A single review; *P < 0.001 (McNemar's paired test). None of the comparisons in the trilobatin batch were statistically significant (P > 0.05).*

3.5 Error Spectrum Analysis

Challenges raised by Model B and subsequently adopted were categorized by type (Table 4). Grading changes comprised 148 downgrades and 58 upgrades. Entry-level counts exported by issue_type included 251 challenges for incorrectly listed downstream effector proteins (215 publications; 474 targets) and 73 for erroneous mappings (68 publications; approximately 85 targets). The "other" category comprised 524 challenges (374 publications), including fabricated evidence types or experimental methods (241), species misclassification (133), and inappropriate use of tool compounds as fallbacks or target-attribution errors (150).

Table 4. Error spectrum of single-model review across the three batches

Error type	Typical manifestation	Corrective mechanism	No. of publications/challenges
Incorrectly listed downstream effector proteins	Cleaved caspase-3, SOD, IL-6, etc. listed as direct targets	Round 2 check (i) + calibration layer	251 challenges
Erroneous mapping	Pseudogenes, non-coding RNAs, or forced isoform mapping	Round 2 check (ii) + standardization module	73 challenges
Overly lenient grading	Purely cellular experiments graded Level II or higher and subsequently downgraded	Round 2 check (iii) + mandatory downgrade	148 publications
Overly stringent grading	Functional-rescue evidence underestimated and subsequently upgraded	Round 2 check (iii) + mandatory upgrade	58 publications
Other	Incorrect inclusion of derivatives or compound formulations, species misclassification, etc.	Initial screening rules + manual final review	524 challenges

3.6 A Representative Case: Correct Retention of the Pathogen Target GP41

PMID 29802645 reported trilobatin as an HIV-1 entry inhibitor targeting the viral envelope glycoprotein GP41, a viral protein rather than a human gene product. Single-model review classified this publication as unverified because GP41 could not be mapped to a human gene. Under the pathogen-target classification rules, GP41 was retained as a pathogen target and assigned Level III.

3.7 Downstream Output: High-Quality Target Library and Core Target Screening for *L. litseifolius*

The three batches yielded a high-quality target library for the active compounds of *L. litseifolius*. A total of 257 traceable records were standardized and consolidated into 74 unique targets (67 human and 7 pathogen targets). Based on the highest evidence level assigned to each component–target pair, 2 were classified as Level Ib, 9 as Level II, 58 as Level III, and 5 as Level IV (Figure 3). Using this library as the positive set, a component–target graph attention network (GAT) was constructed for core target screening, achieving a mean area under the curve (AUC) of 0.9731, an F1 score of 0.8412, and an accuracy of 0.9106 in fivefold cross-validation.

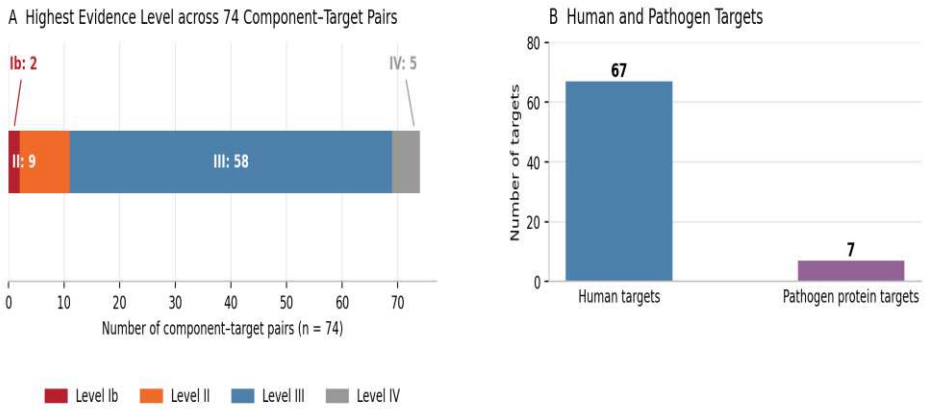


Figure 3. Composition of the high-quality target library for the three active compounds of *L. litseifolius*. (A) Distribution of the highest evidence level for each component–target pair. (B) Composition of human and pathogen targets

4 Discussion

The dual-model adversarial system was developed to address systematic deficiencies in single-LLM literature review. Evaluation against the reference standard showed that, across 776 records, dual-model review increased verification accuracy from 78.5% to 84.8% and five-level grading accuracy from 69.8% to 81.6%, while grading Cohen's κ increased from 0.460 (moderate agreement) to 0.652 (substantial agreement) [19]; both improvements were statistically significant (McNemar's test, $P < 0.001$). The principal value of adversarial review was not wholesale revision of the initial screening results—88.5% of conclusions in the trilobatin batch remained unchanged—but rather the identification of specific errors in challenging publications through structured challenges. This pattern of frequent challenges but relatively infrequent substantive revision is consistent with the design objective and suggests that the marginal benefit of cross-checking is concentrated in difficult publications. Future implementations could use confidence scores or challenge severity to trigger selective cross-checking and reduce computational costs.

Both the paired comparison and the reference-standard evaluation indicated more lenient grading by the single model: 148 records were downgraded compared with 58 upgraded (2.6-fold), and 63 verification decisions changed to unverified compared with 41 to verified. The trilobatin batch showed a similar asymmetry in rule calibration, with 14 mandatory downgrades and 1 mandatory upgrade. This leniency may partly reflect optimistic framing of research conclusions in model training corpora, although this hypothesis requires further investigation. Improvements varied across batches. The larger phloridzin and phloretin batches, which covered more diverse topics and evidence of heterogeneous quality, showed

substantial gains, whereas the two arms performed similarly in the trilobatin batch. The homogeneous topics and regular evidence patterns in the trilobatin literature may explain the smaller incremental benefit in this batch. These findings suggest that cross-checking strategies could be tailored to the topical complexity of the literature.

Because the deterministic layer, comprising consensus merging, calibration, and standardization, does not invoke a language model, the final outputs are reproducible once the model-generated review results have been fixed. This distinguishes the system from pipelines based entirely on LLMs, such as free-form multi-agent debate [20, 21].

Compared with existing component–target databases [22, 23], the targets curated by our system are accompanied by graded evidence and quantitative confidence scores and are traceable to specific PMIDs and experimental methods. In contrast to systematic-review automation tools such as ASReview [24] and RobotReviewer [25], which primarily support inclusion and exclusion decisions, our system grades pharmacological target evidence. Its outputs can therefore directly support component–target knowledge-base construction and downstream computational modeling, as illustrated by the GAT-based core target screening in this study (AUC = 0.9731).

Limitations

This study has several limitations. First, all three batches focused on active compounds of *L. litseifolius*, and generalizability to other species and disease domains requires further evaluation. Second, the reference standard was based on final review by a single researcher; independent review by a second evaluator and assessment of interevaluator agreement were not available, and manual review itself may have introduced bias. Third, the error-spectrum framework was predefined by the system, and classification of compound errors involved subjective judgment; its generalizability across topics remains to be established. Fourth, the system currently covers only English-language literature indexed in PubMed, and extension to Chinese databases such as CNKI and Wanfang Data is planned. Fifth, review was based on titles and abstracts; consequently, key experimental details omitted from abstracts may have led to grading deviations. Finally, one phloridzin publication reverted to the single-model result after an error-checking interface failure, highlighting the need for robust API monitoring and retry mechanisms in production settings.

5 Conclusion

The combination of dual-model adversarial review and deterministic rule calibration significantly improved the accuracy and consistency of literature evidence grading. The system corrected incorrectly listed downstream effector proteins, erroneous mappings, and grading deviations produced by single-model review and generated component–target records with traceable evidence and quantitative grades, supporting downstream knowledge-base construction and computational modeling. This framework offers a reusable approach to scalable, high-quality biomedical knowledge-base construction and evidence-based data governance.

References

[1] Zhai Y, Liu L, Zhang F, et al. Network pharmacology: a crucial approach in traditional Chinese medicine research[J]. Chinese Medicine, 2025,20:8.

[2] Yu Z, Wu Z, Wang Z, et al. Network-Based Methods and Their Applications in Drug Discovery[J]. Journal of Chemical Information and Modeling, 2023,64(1):57-75.

[3] Duan XC, Huang S, Peng DY, et al. Application of network pharmacology in research on traditional Chinese medicine compound formulas[J]. Chinese Pharmacological Bulletin, 2020,36(3):303-308. (in Chinese)

[4] Zhao X, Chen J, Zhang H, et al. Insights into Advantages and Limitations of Network Pharmacology in Traditional Chinese Medicine[J]. Future Integrative Medicine, 2025,4(4):255-256.

[5] Thombre K R, Gabhane V, Gupta K R, et al. Network pharmacology in the multi-omics era: uncovering novel therapeutic strategies[J]. Discover Chemistry, 2026,3:111.

[6] Yang L, Wang H, Zhu Z, et al. Network Pharmacology-Driven Sustainability: AI and Multi-Omics Synergy for Drug Discovery in Traditional Chinese Medicine[J]. Pharmaceuticals, 2025,18(7):1074.

[7] Fang S, Dong L, Liu L, et al. HERB: a high-throughput experiment- and reference-guided database of traditional Chinese medicine[J]. Nucleic Acids Research, 2020,49(D1):D1197-D1206.

[8] Guo E, Gupta M, Deng J, et al. Automated Paper Screening for Clinical Reviews Using Large Language Models: Data Analysis Study[J]. Journal of Medical Internet Research, 2024,26:e48996.

[9] Yang Y, Ma Y, Feng H, et al. Minimizing Hallucinations and Communication Costs: Adversarial Debate and Voting Mechanisms in LLM-Based Multi-Agents[J]. Applied Sciences, 2025,15(7):3676.

[10] Bruford E A, Braschi B, Denny P, et al. Guidelines for human gene nomenclature[J]. Nat Genet, 2020,52(8):754-758.

[11] Zhao P, Yang Z, Tan Z, et al. Leaf color as an indicator of the accumulation pattern of pharmaco-nutritional compounds in Lithocarpus litseifolius spring foliage[J]. Frontiers in plant science, 2025,16:1726934.

[12] Sun L, Guan R, Gao M, et al. Myocardial ischemia–reperfusion injury after acute myocardial infarction: spatiotemporal mechanisms and endotype-matched multi-target modulation by compounds from traditional Chinese medicine[J]. Frontiers in Cardiovascular Medicine, 2026,13(1):1797330.

[13] Yang B, Olsen M, Vali Y, et al. Study designs for comparative diagnostic test accuracy: A methodological review and classification scheme[J]. Journal of Clinical Epidemiology, 2021,138(1):128-138.

[14] Cohen J. A Coefficient of Agreement for Nominal Scales[J]. Educational and Psychological Measurement, 1960,20(1):37-46.

[15] Branco D F, Botti P, Bouredoucen H, et al. Diagnostic Accuracy of Artificial Intelligence Versus Musculoskeletal Radiologists for Foot and Ankle Fracture Detection on Radiographs[J]. Diagnostics, 2026,16(16):2507.

[16] McNemar Q. Note on the sampling error of the difference between correlated proportions or percentages[J]. Psychometrika, 1947,12(2):153-157.

[17] Ahn S. Building and analyzing machine learning-based warfarin dose prediction models using scikit-learn[J]. Translational and Clinical Pharmacology, 2022,30(4):172.

[18] Zeng M, Li Y, Zhu Y, et al. Inflammatory markers and clinical factors as key independent risk factors for frailty: a retrospective study[J]. BMC Geriatrics, 2025,25(1):404.

[19] Landis J R, Koch G G. The Measurement of Observer Agreement for Categorical Data[J]. Biometrics, 1977,33(1):159.

[20] Delgado-Chaves F M, Jennings M J, Atalaia A, et al. Transforming literature screening: The emerging role of large language models in systematic reviews[J]. Proceedings of the National Academy of Sciences, 2025,122(2):e1883005174.

[21] Oami T, Okada Y, Nakada T. Performance of a Large Language Model in Screening Citations[J]. JAMA network open, 2024,7(7):e2420496.

[22] Wang Y, Liu M, Jafari M, et al. A critical assessment of Traditional Chinese Medicine databases as a source for drug discovery[J]. Frontiers in Pharmacology, 2024,15:1303693.

[23] Ren X, Yan C, Zhai R, et al. Comprehensive survey of target prediction web servers for Traditional Chinese Medicine[J]. Heliyon, 2023,9(8):e19151.

[24] van de Schoot R, de Bruin J, Schram R, et al. An open source machine learning framework for efficient and transparent systematic reviews[J]. Nature machine intelligence, 2021,3(2):125-133.

[25] Marshall I J, Kuiper J, Wallace B C. RobotReviewer: evaluation of a system for automatically assessing bias in clinical trials[J]. Journal of the American Medical Informatics Association : JAMIA, 2016,23(1):193-201.