MARS: A Multi-Agent Framework for Reviewing Observing Proposals

An Automated Evidence-Grounded, Multi-Perspective Assistant for Reviewing Astronomical Observing Proposals for Time Allocation/Observation Committees

F. AIT BENKHALI, F. JANKOWSKY, S. WAGNER

Abstract

We present MARS (Multi-Agent Review System), an advisory AI framework that generates structured, evidence-grounded draft assessments for reducing the substantial workload of human reviewers. Its five-role pipeline combines three parallel specialist analyses – scientific merit, technical feasibility, and methodology/statistic – with adversarial critique and consensus synthesis. Substantive findings are linked to typed evidence, including proposal excerpts, verified literature records, and deterministic calculator outputs. Unsupported claims are explicitly flagged. A facility-configurable evaluation rubric is applied through a deterministic scoring layer separated from generative model judgments, while uncertainty and unresolved disagreements remain visible. MARS supports, but does not replace, expert judgment: It computes structured assessments and rubric scores. Human reviewers retain full authority to inspect, revise, and finalize assessments.

Motivation & Challenges

Reviewers of astronomical proposals need to assess scientific merit, literature context, technical feasibility, observation times, and statistical robustness – while maintaining consistency and transparency. AI assisted review is promising, but generating a convincing review is not the same as producing a reliable one:

Potential benefits: AI can support structured assessments, identify weaknesses, assist with first order review

Evaluation bias: LLM-based judges exhibit position, verbosity, and self-preference biases.

Security risks: AI agents using external tools can be vulnerable to malicious instructions in untrusted content.

Coordination failures: AI systems can share blind spots, converge prematurely, or produce inconsistent findings

Challenge: Turning promising AI capabilities into a reliable, secure, and auditable scientific review process requires explicit verification, uncertainty management, and human oversight.

How MARS Addresses These Challenges

How can AI-generated scientific assessments be made evidence-grounded, auditable, secure, and useful without displacing expert judgment? MARS addresses this engineering challenge through five design objectives :

1. **Evidence before assertion:** Link substantive findings to inspectable proposal passages, verified literature records, or deterministic calculations; explicitly flag unverified claims.
2. **Structured independent analysis:** Combine three specialized reviewers with adversarial critique and synthesis that preserves unresolved disagreements.
3. **Deterministic verification:** Separate numerical calculations and rule-based score aggregation from generative model judgments.
4. **Auditability and security:** Record computational provenance, support replay and verification, and protect the pipeline against untrusted inputs and unsafe tool execution.
5. **Human authority:** Produce advisory draft reviews for expert inspection, correction, and final decision-making not autonomous allocation decisions.

MARS shifts the focus from generating reviews to engineering a review process whose evidence, uncertainties, and computational history can be inspected.

Evidence-Grounded Claims

Every substantive claim is linked to typed evidence with an explicit verification status. Unverifiable claims are flagged rather than presented as established facts

Proposal excerpts: Exact text spans linked to the normalized proposal.

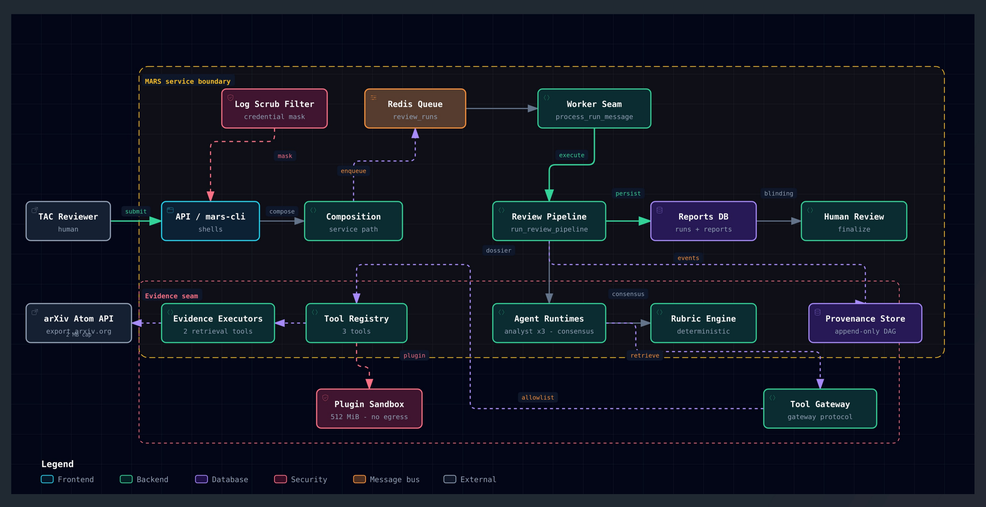
Literature records: Bibliographic metadata and source quotations verified against the original abstract.

Calculator outputs: Reproducible numerical checks of visibility, sensitivity, and exposure time.

Unverified inferences: Explicitly flagged assumptions or claims lacking verified support.

Conflicting evidence: Contradictory findings and their sources preserved for human review.

Verification boundary: Verifying a quotation does not verify its scientific interpretation. Evidence grounding improves traceability but does not guarantee correctness.



Hierarchical Agent Workflow

Agent/Role	Main responsibility	Typical output
Scientific Reviewer	Assess novelty, theoretical soundness, and broader scientific impact	Evidence-linked strengths, weaknesses, and criterion scores
Technical Reviewer	Examine observational feasibility, target visibility, instrument constraints, sensitivity & exposure time	Feasibility findings and calculator records
Methodology Reviewer	Examine analysis readiness, statistical assumptions, systematic uncertainties, and robustness	Findings and calculator record
Adversarial Critic	Cross-check all three reports for contradictions, unsupported claims, and overconfidence	Critical report; no independent numerical score
Consensus Builder	Combine findings, respond to critique, and retain unresolved disagreement	Structured consensus report and advisory draft

Scoring, Auditability & Human Oversight

MARS combines structured rubric scoring, explicit uncertainty reporting, traceable execution, and human control.

Explicitly six-criterion rubric: Scientific novelty, theoretical soundness, observational feasibility, sensitivity adequacy, data-analysis readiness, and broader scientific impact.

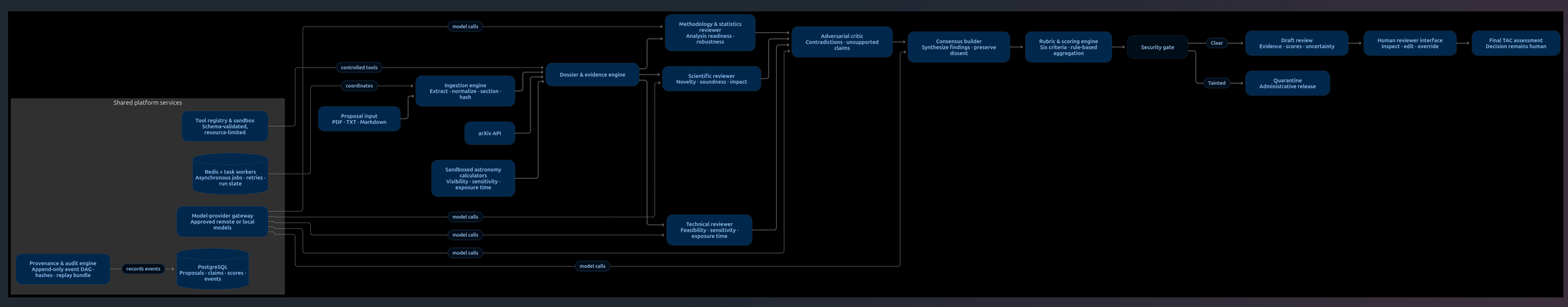
Transparent scoring: Criterion-level scores and justifications are derived from structured agent findings using explicit rubric rules.

Visible uncertainty: Reports expose evidence links, unverified claims, missing information, critic warnings, unresolved disagreements, and measured run-to-run variability.

Human oversight: Reviewers can inspect evidence, edit scores and narratives, record overrides, and optionally assess proposals before seeing AI-generated drafts.

Auditable provenance & replay: Hash-linked records capture pipeline events, model calls, tool executions, scores, and human overrides. Portable run bundles support inspection and replay; discrepancies are flagged.

Security & failure visibility: Prompt-injection checks, restricted tools, sandboxing, and quarantine mitigate risks. Unverifiable citations, tool failures, and ambiguous calculations trigger explicit warnings.



Gefördert durch:

Bundesministerium für Forschung, Technologie und Raumfahrt

www.deutscheszentrumastrophysik.de

DEUTSCHES ZENTRUM ASTROPHYSIK