Skepthical review: *The Parallel Science Project: Cyber Space for Human–AI Co-Evolution of Science*

Summary

The paper proposes Parallel Science, an open infrastructure intended to host and curate a dedicated AI-generated scientific literature that runs in parallel to (but interoperates with) the human literature. The concrete implementation described centers on Parallel ArXiv (papers.parallelscience.org), which ingests AI-generated papers from GitHub Pages, assigns stable PX identifiers (PX:YYMM.NNNNN), versions papers using content hashing, and classifies them into an arXiv-compatible taxonomy via a two-step LLM-based classifier (Secs. 3–5). The manuscript also outlines a production stack of supervised and autonomous AI "scientist" fleets (Denario/CosmoEvolve) with containerized deployments across multiple providers plus monitoring/cost accounting ("Mission Control"), and introduces Parallel Open Review (reviews.parallelscience.org) as an AI review layer envisioned to connect to replication engines and drive a production–evaluation–selection loop (Secs. 4, 7–8). Overall, the systems/institutional framing is timely and the Parallel ArXiv mechanics are described with useful implementation detail, but the review/replication loop, governance/safety posture, security/threat model, and quantitative evaluation of system behavior are currently under-specified relative to the central claims, making it hard to assess robustness, quality control effectiveness, and real-world scalability.

Strengths

- Clear institutional motivation for separating AI-generated outputs from the human scholarly record while enabling controlled interaction (Secs. 2.1–2.2).
- Concrete, end-to-end description of Parallel ArXiv: PX identifiers, versioning logic, hashing, ingestion from GitHub Pages, storage/DB choices, and planned API (Secs. 3.1–3.7, 5).
- Strong emphasis on provenance and auditability via Git history, immutable identifiers, and explicit versioning (Secs. 3.2, 5, 8.1).
- Pragmatic, decoupled publication interface (scraping a template/metadata from GitHub Pages) that could allow multiple upstream AI-scientist systems to publish into the same repository (Secs. 3.5, 5).
- Detailed fleet operations description (multi-provider LLM usage, containerization, supervised vs autonomous modes), suggesting operational experience rather than purely conceptual design (Secs. 4.1–4.4).
- Cost transparency is treated as a first-class feature (Mission Control and per-paper cost reporting), which is often missing in related systems (Sec. 7).
- The paper is relatively candid about limitations and ethical concerns (pollution/attribution/energy), and the examples illustrate feasibility of nontrivial workflows at low LLM API cost (Secs. 6–8.5).

- Figure 1 usefully communicates the intended workflow and the separation of system components at a high level.

Major issues

1. **The central production–evaluation–selection claim is not yet matched by an equally concrete description and evaluation of Parallel Open Review, the replication engine, and the compute/resource allocation policy.** These components are positioned as the key scaling mechanism and quality-control backstop, but their current capabilities vs. planned features are unclear, and there is little detail on review schemas, model configuration/calibration, linkage to paper versions, replication frequency, or how review/replication signals actually reallocate resources across fleets (Secs. 1, 5, 7, 8.2).

   Recommendation: Expand Sec. 7 and Sec. 8.2 (and align claims in Sec. 1 and Sec. 5) with a clear "system status" breakdown: (1) a table listing what is deployed now vs. in development vs. aspirational for Parallel Open Review and any replication components; (2) the exact review template/schema (sections, scores, confidence, checklists), reviewer model(s), prompting strategy, aggregation, and how reviews are versioned and linked to specific PX paper versions; (3) any existing calibration/validation (even small-scale), e.g., agreement between multiple AI reviewers and/or spot-checked correlation with human judgments; (4) the concrete resource allocation mechanism (metrics, update cadence, thresholds, and how it changes task queues/compute budgets). If the closed loop is not yet operational, explicitly reframe it as future work and narrow the paper's claims accordingly.

2. **The paper lacks a systematic quantitative evaluation of infrastructure behavior and quality-control effectiveness.** Beyond a few case studies and a small cost table, there is no measurement of ingestion robustness, classification accuracy/stability, review reliability, throughput, failure rates, or recovery behavior. This limits the ability to assess scalability, reliability, and whether the proposed safeguards can work in the high-volume regime the paper motivates (Secs. 3.3, 3.5, 4–5, 7–8).

   Recommendation: Add an evaluation section (likely in Sec. 7–8) reporting operational metrics from the running system, e.g.: (1) Parallel ArXiv ingestion latency distribution, webhook/cron failure rates, and recovery success (Secs. 3.5, 5); (2) PX allocation/registry integrity metrics (collision/duplicate checks, missing-entry checks) and any load/burst tests (Secs. 3.2–3.5); (3) LLM category classification accuracy on a manually labeled audit set, plus stability across model updates (Sec. 3.3); (4) Parallel Open Review statistics (coverage rate, median cost per review, score distributions, inter-reviewer agreement) and—if possible—small-scale human comparison (Secs. 7, 8.2); (5) fleet productivity metrics (papers/week, mean time-to-publish, fraction requiring human intervention, tool/LLM failure rates) (Secs. 4–5). Even modest-scale measurement would significantly strengthen the paper's core systems contribution.

3. **Governance, accountability, and moderation mechanisms remain high-level given that avoiding pollution, handling harmful content, and maintaining trust are core motivations.** The paper does not specify concrete policies for submission eligibility/curation, spam and duplication control, retractions/corrections, appeals, human moderation, or how accountability is assigned when papers are authored by systems. It also under-discusses negative systemic dynamics (over-optimization to AI reviewer metrics, training feedback loops/model collapse, homogenization of research directions) (Secs. 2.1–2.2, 8.1, 8.4–8.5).

   Recommendation: Add a dedicated governance/safety subsection (extend Sec. 2.1 and Sec. 8.4–8.5) that specifies: (1) eligibility criteria for pipelines/systems publishing to Parallel ArXiv, and any rate limits/quotas; (2) moderation workflow (AI/human flagging, takedown/retraction/correction policy, audit trail, and how "immutability of record" coexists with legal/ethical redaction); (3) spam/duplication detection and minimum quality gates; (4) dual-use/harmful-content handling and red-teaming plans; (5) accountability statement (who is responsible for corrections/harms) and how human supervision is recorded; (6) monitoring/mitigations for feedback loops and selection-induced homogenization (e.g., periodic corpus audits, reviewer diversification, human oversight triggers). Where not implemented, label as open problems with concrete next steps (e.g., community oversight board).

4. **Security and abuse-resilience are underdeveloped for a system that scrapes untrusted HTML/PDF content and uses LLMs for classification/review.** The manuscript mentions webhook HMAC validation but does not provide a threat model or defenses against prompt injection (aimed at classifier/reviewer), malicious HTML/PDF payloads, DoS/bursty submissions, repository takeover within the GitHub org, or sandboxing of parsers/renderers (Secs. 3.3, 3.5, 5).

   Recommendation: In Sec. 5 (and where relevant Sec. 3.3/3.5), add an explicit threat model and mitigations: (1) sandboxing/isolating HTML/PDF fetching and parsing; (2) file-type validation, size limits, rate limiting, and queue backpressure; (3) prompt-injection defenses (content sanitization, instruction hierarchy, classifier inputs restricted to metadata/abstract, ensemble voting, adversarial tests); (4) access control within the GitHub org (branch protections, required reviews, signed commits) and publication authorization model; (5) monitoring/alerting for anomalous publishing patterns. Include current limitations and planned hardening steps.

5. **Reproducibility is emphasized via Git provenance and hashing, but operational reproducibility guarantees and artifacts are not clearly specified.** It is unclear whether the example PX papers have fully public repos with runnable code/data, whether Docker images and package versions are pinned, how model versions/prompts are logged, how nondeterminism from LLM calls is handled, and how secrets/credentials are managed for reproducible deployments (Secs. 5–6, 8.1).

Recommendation: In Sec. 5 and Sec. 8.1, provide an explicit reproducibility checklist and confirm what is publicly available for PX:2604.00016, PX:2604.00009, and PX:2604.00015: (1) repository links, exact commit SHAs, Dockerfiles/lockfiles, and environment capture; (2) datasets (or hashes/IDs and download scripts), random seeds, and run scripts (ideally "one-command reproduce"); (3) logging policy for prompts, model identifiers/versions, tool calls, and (where feasible) responses; (4) handling of nondeterminism (e.g., replay logs vs rerun expectations); (5) secrets management and support for private artifacts. This can be concise but should be concrete.

6. **Versioning semantics and hashing appear to be based primarily on metadata fields (title/authors/abstract/categories), which can miss substantive PDF/body changes or trigger versions for minor metadata edits.** This risks undermining the stated goal of an immutable, auditable scientific record and complicates replication/review linkage (Secs. 3.2, 3.5).

   Recommendation: Revise Sec. 3.2 to define what constitutes a "new version" and update the hashing/versioning strategy accordingly. Consider including (1) a normalized PDF hash and/or (2) the source repository commit hash (or a manifest hash over source files) in addition to metadata. Clearly describe how reviews/replications bind to a specific version and what happens when only metadata changes. If the current system intentionally versions on metadata-only, justify the trade-off and document expected failure modes.

7. **The scope and contribution relative to Denario and CosmoEvolve are not sufficiently disentangled.** These pipelines are repeatedly referenced as central "AI scientist" systems producing the example outputs, but their methodological details are largely deferred, making it hard to judge which parts are contributions of this paper (Parallel Science infrastructure) vs. separate research-agent work (Secs. 1, 4.1, 4.4, 6.1–6.3, 8.2).

   Recommendation: Clarify scope in Sec. 1 and Sec. 4: (1) add a concise summary of Denario extensions beyond prior work (e.g., iterative refinement loops, supervision mechanics, self-healing publishing), and likewise summarize CosmoEvolve's loop at a level sufficient to understand how it interfaces with Parallel Science; (2) in Sec. 6, explicitly label the examples as infrastructure feasibility demonstrations vs. scientific-method contributions, and state what evidence they provide for the paper's claims; (3) link to public docs/repos where deeper pipeline details live, if not in this manuscript.

8. **The interaction model between Parallel ArXiv and the human literature is compelling but operationally vague.** The paper discusses "porous" interaction and cross-citation, but does not specify concrete citation formats/BibTeX conventions that preserve provenance, discoverability strategies (cross-indexing/search bridges), or

how fragmentation of citation networks will be mitigated. Accountability and provenance labeling for downstream human use are central to the institutional argument (Secs. 2.2, 3.6, 8.3, 9).

Recommendation: Operationalize the "porous boundary" in Sec. 2.2, Sec. 3.6, and Sec. 8.3 by adding: (1) a recommended citation format + BibTeX entry that clearly marks AI provenance and PX identifiers; (2) UX/search plans to keep AI-origin visible while enabling discovery; (3) concrete cross-indexing/API plans (e.g., mirror endpoints, semantic search bridging) and rate limits/terms; (4) a worked example of a human paper citing a PX paper and how that should be interpreted. Also briefly discuss alternative designs (e.g., AI-tagging within arXiv) and justify the separate repository choice with mitigation strategies.

Minor issues

1. The LLM-based two-stage category classification is central to navigation but remains under-specified and unevaluated: exact models, prompts, token limits, confidence/abstention, retries, and handling of adversarial text are not described in sufficient detail (Sec. 3.3).

   Recommendation: Expand Sec. 3.3 with a compact implementation table: models/versions, prompt format (and whether examples are used), input fields, token limits, validation rules, retry/abstention policy, and how invalid categories are handled. Add at least a small manual audit (accuracy/confusions) and note defenses against prompt injection.

2. Identifier allocation and ingestion robustness under concurrency/high load are not discussed in enough operational detail. The registry design (SQLite + GCS syncing) raises questions about atomicity, single-writer assumptions, and recovery from partial failures or bursty webhook events (Secs. 3.2–3.5).

   Recommendation: In Sec. 3.2–3.5, describe concurrency control and failure recovery: locking/transaction strategy, deduplication of webhook events, idempotency, monitoring for collisions/missing IDs, and what guarantees hold under burst load. If the system relies on a single writer, state it explicitly and outline a scaling plan.

3. The ingestion interface is described as upstream-agnostic, but the required HTML/metadata schema for GitHub Pages scraping is only informally mentioned, limiting external adoption (Secs. 3.5, 5).

   Recommendation: Specify the exact expected metadata schema (meta tags / JSON-LD) and provide a minimal example snippet. Link to a template repository or schema definition so external systems can integrate without reverse-engineering.

4. Mission Control and cost transparency are highlighted but supported by limited monitoring validation and limited aggregate statistics beyond three example papers (Sec. 7).

   Recommendation: Augment Sec. 7 with summary stats over more runs (per-paper cost distribution, variance, utilization, failure rates), and briefly describe how monitoring correctness is validated. Clarify what dashboards/logs are public vs internal.

5. Fleet operations discuss tools/providers but omit robustness and safety mechanisms (timeouts are mentioned, but not fallbacks, provider outages, model routing, runaway-cost prevention, or tool-use guardrails) (Secs. 4.1–4.2).

   Recommendation: Add a short subsection in Sec. 4.1–4.2 describing retries/backoff, multi-provider fallbacks, model selection/routing heuristics, guardrails for tool use, and policies to stop runaway loops/cost explosions. If available, include empirical observations from operations.

6. Example outputs in Sec. 6 include performance metrics but are not clearly positioned (feasibility demo vs. competitive scientific result), and lack baseline comparisons or limitations, which can miscalibrate readers about what the infrastructure demonstrates (Secs. 6.1–6.3, 8.4).

   Recommendation: In Sec. 6.1–6.3, add brief baselines or literature anchors where feasible, and explicitly label these as preliminary feasibility demonstrations if they are not intended to be state of the art. Tie the positioning to Sec. 8.4 limitations.

7. The institutional argument for a separate Parallel ArXiv would benefit from a more explicit comparison to alternatives (e.g., AI provenance tags inside existing repositories) and a clearer discussion of discoverability/citation fragmentation trade-offs (Secs. 2.1–2.2).

   Recommendation: Add a concise alternatives-and-tradeoffs paragraph in Sec. 2.1–2.2 and explain mitigation plans (cross-indexing, unified search, explicit provenance in citations).

8. Figure 1 has clarity issues: missing/ambiguous arrow directionality and legend; inconsistent depiction of LLM providers; and the feedback loop from review/replication to resource allocation is not explicit despite being claimed in the caption (Fig. 1; Secs. 7–8.2).

   Recommendation: Revise Fig. 1 to add: (1) a labeled resource allocator and explicit feedback arrows from reviews/replications; (2) consistent connections to LLM providers (or a shared service box); (3) arrowheads and a legend for line styles/colors; (4) clearer depiction of which systems push to GitHub repos vs GitHub Pages.

9. Platform dependence and portability are under-discussed: reliance on GitHub Pages/GitHub org workflows, specific clouds, and proprietary LLM APIs may affect long-term stability and equitable access (Secs. 3.5, 4, 8.5).

Recommendation: In Sec. 8.4–8.5, explicitly discuss portability plans (alternative publication backends beyond GitHub Pages, self-hosting options, abstraction of LLM providers) and the implications for equity of access and sustainability.

Very minor issues

1. Minor formatting/typographical inconsistencies (stray Markdown markers like "###"/"##", spacing around math and hyphenated terms, inconsistent PX identifier punctuation) reduce polish (Secs. 2–4, 6, 9).

   Recommendation: Proofread and normalize headings/Markdown artifacts, spacing around math/hyphenation (e.g., "denario-i fleet"), and apply a consistent PX identifier style throughout text and figures.

2. Bibliography and in-text citations are inconsistent (organization authors, capitalization, incomplete entries, occasional mismatch between in-text citation and reference list) (Sec. 9, References).

   Recommendation: Standardize references to a single style with consistent author/organization names, years, titles, venues, and URLs; ensure all in-text citations match reference entries.

3. Inconsistent terminology/capitalization and some acronyms used before definition (e.g., arXiv vs ArXiv, Denario naming, LaTeX capitalization, VLM/EDA/PX) slightly harm clarity (Secs. 3–6).

   Recommendation: Perform a copy-edit pass to standardize naming and define acronyms at first use (including in table/figure captions).

4. Some URLs/hostnames appear inconsistently formatted and may not be clickable; Fig. 1 likely contains a typo in the Parallel Open Review URL label (Abstract, Fig. 1, Sec. 9).

   Recommendation: Standardize URL formatting/casing, ensure clickable links in the final PDF (\url/\href), and correct any hostname typos in figures and captions.

5. Ambiguous grid-size notation "1283 periodic grid" is likely intended as 128^3 (Sec. 6.1).

   Recommendation: Rewrite as "128^3" or "$128 \times 128 \times 128$" to remove ambiguity.

6. Statistical notation is inconsistent (e.g., "R2" instead of R^2) (Sec. 6.1).

   Recommendation: Typeset coefficient of determination consistently as R^2.

Key statements and references

- ✔ **Large language models and multi-agent systems are now capable of generating scientific hypotheses, writing and executing code, analyzing data, and producing coherent research papers, as demonstrated in recent work on fully automated scientific discovery, autonomous chemical research, and benchmarking language agents on machine learning experimentation.**
- *Reference(s):* Lu et al., 2024, Boiko et al., 2023, Huang et al., 2024
- *Justification:* Verification failed with gpt-5: Error code: 400 - {'error': {'message': 'Your input exceeds the context window of this model. Please adjust your input and try again.', 'type': 'invalid_request_error', 'param': 'input', 'code': 'context_length_exceeded'}}
- ⚠ **The research backend CMBAgent implements a two-phase planning-and-control strategy in which a planner agent designs an execution plan and a controller agent delegates subtasks to specialized agents, and is built on prior work on multi-agent systems for cosmological parameter analysis and open-source planning and control systems with language agents for autonomous scientific discovery, orchestrated using AG2 for multi-agent conversation patterns and LangGraph for structured generation tasks.**
- *Reference(s):* Laverick et al., 2024, Xu et al., 2025, Wu et al., 2023
- *Justification:* Xu et al., 2025 describes cmbagent as a two-phase Planning & Control system: a planner designs a multi-step plan and, in the Control phase, a controller delegates each sub-task to specialized agents such as the researcher and engineer (Xu et al., 2025). It is powered by AG2/autogen (Xu et al., 2025; Laverick et al., 2024), and builds on earlier MAS work for cosmological parameter analysis (Laverick et al., 2024). However, while LangGraph is used in the denario system to turn results into manuscripts (a separate subsystem where cmbagent is the backend), the papers do not show cmbagent itself being orchestrated with LangGraph (Xu et al., 2025). Thus the claim about LangGraph within cmbagent is not directly supported, making the statement only partially supported.
- ✔ **The Denario research pipeline used in this infrastructure is based on the Denario project framework for deep knowledge AI agents for scientific discovery, originally described by Villaescusa-Navarro et al. (2025), which introduced a dual-backend architecture that the present work extends with iterative refinement across multiple research cycles and fleet-scale containerized deployment.**
- *Reference(s):* Villaescusa-Navarro et al., 2025
- *Justification:* Verification failed with gpt-5: Error code: 400 - {'error': {'message': 'Failed to download file. File urls cannot be larger than 32MB.', 'type': 'invalid_request_error', 'param': 'url', 'code': 'file_above_max_size'}}

- ✖ **The CosmoEvolve Virtual Lab system, which operates through the same CMBAgent stack and OpenClaw gateway described here, builds on prior work on multi-agent systems for cosmological parameter analysis to perform autonomous cosmological data analysis and will be described in detail in a forthcoming publication referenced as a multi-agent system for cosmological parameter analysis (2024).**
- *Reference(s):* Laverick et al., 2024
- *Justification:* Laverick et al., 2024 describes the cmbagent/CMBAgent multi-agent system and its architecture but does not mention any 'CosmoEvolve Virtual Lab' or an 'OpenClaw gateway.' It also states the current system requires human feedback at every step rather than being autonomous. No forthcoming publication about a CosmoEvolve system is referenced. Thus, the statement is not supported by the paper.

Mathematical consistency audit

This section audits **symbolic/analytic** mathematical consistency (algebra, derivations, dimensional/unit checks, definition consistency).

Maths relevance: light

The paper is primarily a systems/infrastructure description with very limited mathematics: identifier-format notation, hashing/versioning rules, occasional complexity notation ($O(1)$), and brief summaries of external example papers that mention standard statistical/ML quantities (PCA variance explained, R^2, negative log-likelihood minimization). There are no displayed equations or derivations to audit for algebraic correctness.

Checked items

1. ✔ **PX identifier format and example consistency** (Sec. 3.2, p.3 (parsed pages 3/12))
    - **Claim:** Identifiers have the format PX:YYMM.NNNNN, with an example PX:2604.00001 described as the first paper indexed in April 2026.
    - **Checks:** notation consistency, constraint/sanity check
    - **Verdict:** PASS; confidence: high; impact: minor
    - **Assumptions/inputs:** YYMM corresponds to year (two digits) and month (two digits)., NNNNN is a zero-padded sequence number within that month.
    - **Notes:** The example PX:2604.00001 matches April 2026 under the stated YYMM convention, and the prefix PX is used consistently as the Parallel ArXiv distinguisher.
2. ✔ **Append-only registry + sequence table logic** (Sec. 3.2, p.3 (parsed pages 3/12))

- **Claim:** An append-only registry maps repository names to identifiers, and an ID sequence table provides $O(1)$ allocation of the next available number within each month.
- **Checks:** logic consistency, asymptotic-claim plausibility (non-numeric)
- **Verdict:** PASS; confidence: medium; impact: minor
- **Assumptions/inputs:** A per-(year,month) counter/sequence is maintained in the database., Allocation increments the stored counter and returns the previous/new value deterministically.
- **Notes:** No internal contradictions: an explicit per-month sequence record can yield constant-time lookup/update. Formal proof is not provided, but the claim is consistent with the described data structure.

3. ✔ **Content-hash versioning rule consistency** (Sec. 3.2, p.3 (parsed pages 3/12))
    - **Claim:** A SHA-256 hash of title, author, abstract, and categories is used to detect content changes; hash changes trigger a new version and prevent spurious version bumps when rebuilding pages without content changes.
    - **Checks:** definition consistency, logic/sanity check
    - **Verdict:** PASS; confidence: high; impact: minor
    - **Assumptions/inputs:** Hash input fields are exactly those listed (title, author, abstract, categories)., Rebuilds that do not alter these fields should keep the hash stable.
    - **Notes:** Given the stated hashed fields, rebuild-only changes (e.g., HTML regeneration) would not affect the hash, matching the stated goal. The rule is self-consistent as described.

4. ⚠ **Turbulent-fluid example: grid-size notation and method chain** (Sec. 6.1, p.8 (parsed pages 8/12))
    - **Claim:** Example paper uses density/velocity fields on a "1283" periodic grid over 10 time slices; computes spatial derivatives spectrally and temporal derivatives via finite differences; builds a 66-term library; applies cross-validated LASSO with OLS refinement; reports R^2 and stability under integration.
    - **Checks:** notation clarity, workflow logical consistency
    - **Verdict:** UNCERTAIN; confidence: medium; impact: minor
    - **Assumptions/inputs:** "1283" is intended to mean a 3D grid with 128 points per dimension (128^3)., The pipeline description is a narrative summary rather than a derivation.
    - **Notes:** The chain of methods is narratively coherent, but the symbol-like token "1283" is ambiguous (128^3 vs 1,283). This ambiguity prevents a clean symbolic interpretation, though it does not create a direct contradiction elsewhere in this paper.

5. ⚠ **VCSF anomaly score definition (min over parameter space)** (Sec. 6.2, p.8 (parsed pages 8/12))
    - **Claim:** Anomaly scores are computed as the minimum negative log-likelihood over the parameter space, found via gradient-based optimization, after PCA and whitening against nuisance parameters and modeling with a conditional normalizing flow.
    - **Checks:** definition consistency, missing-symbol audit
    - **Verdict:** UNCERTAIN; confidence: low; impact: minor
    - **Assumptions/inputs:** A conditional likelihood $p(x|\theta)$ (or similar) is defined by the flow, and the anomaly score is $s(x) = \min_\theta[-\log p(x|\theta)]$., The "parameter space" refers to the nuisance/conditioning variables referenced earlier in the paragraph.
    - **Notes:** The minimization-based score is plausible and internally consistent at the narrative level, but the paper does not introduce symbols/domains (what is optimized over, constraints, differentiability assumptions) needed to verify formal correctness of the objective and its relationship to "invariance to known physical parameter variations."

Limitations

- The manuscript contains no displayed equations, numbered formulas, or step-by-step derivations; therefore an algebraic/derivational consistency audit is largely inapplicable.
- Sections 6.1–6.3 summarize other papers' methodologies and metrics without providing the underlying definitions/equations in this PDF, so only surface-level notation/logic checks are possible.
- No unit/dimensional analysis is feasible because the paper does not define physical quantities symbolically (it discusses infrastructure rather than presenting physical models).

Numerical results audit

This section audits **numerical/empirical** consistency: reported metrics, experimental design, baseline comparisons, statistical evidence, leakage risks, and reproducibility.

Of 20 candidate numerical/consistency checks, 15 passed and 5 were uncertain; no failures were detected. Verified items include exact parsing/format checks (identifier example), exact arithmetic/unit conversions (seconds↔minutes; polling period), exact integer recomputations (128^3 grid and totals), simple inequality/bounds checks (R^2 range, percentage and metric bounds, σ threshold comparison, frequency ordering), and a combinatorial parts-to-total index coverage check for $N = 12$ in Table 1. Uncertain items require additional document text/table extraction or external quantities (e.g., DB size vs paper count; document-wide repeated-constant scans; cross-table identifier matching).

Checked items

1. ✔ **C1** (Page 3, Section 3.2 (Identifier scheme and versioning))
    - **Claim:** Identifier format PX:YYMM.NNNNN; example: "PX:2604.00001 for the first paper indexed in April 2026".
    - **Checks:** format_consistency (example parsing vs claimed meaning)
    - **Verdict:** PASS
    - **Notes:** Parsed YYMM and sequence compared to claimed meaning.
2. ✔ **C2** (Page 3, Section 3.3 (Category classification))
    - **Claim:** Classification selects one of "21 top-level archives" and assigns "up to three" secondary categories for cross-listing.
    - **Checks:** integer_reasonableness / bounds check from stated limits
    - **Verdict:** PASS
    - **Notes:** Constraint constants checked; schema-level validation only.
3. ⚠ **C3** (Page 4, Section 3.4 (Database and storage))
    - **Claim:** “The database is compact, at approximately 2 KB per paper.”
    - **Checks:** cheap_recomputation (order-of-magnitude check given paper counts if available)
    - **Verdict:** UNCERTAIN
    - **Notes:** Need number_of_papers and/or total DB size to recompute per-paper storage.
4. ✔ **C4** (Page 6, Table 1 caption + rows (Supervised fleet configuration))
    - **Claim:** Table 1 describes supervised denario-i fleet with $N = 12$ and lists scientist indices $1, 4, 5$; 2; 3; 6; 7–12.
    - **Checks:** parts_to_total (index coverage equals N)
    - **Verdict:** PASS
    - **Notes:** Checked that listed indices/range cover exactly $1..N$ with no omissions/extras.
5. ✔ **C5** (Page 6, Table 1 (Scientist 3 row))
    - **Claim:** Scientist 3 has "CPU 32", "RAM 64 GB", "GPU RTX PRO 6000", "Timeout 1800 s".
    - **Checks:** unit_consistency (time conversion) + sanity ratio
    - **Verdict:** PASS
    - **Notes:** Converted seconds to minutes and compared.
6. ✔ **C6** (Page 6, Table 1 (most rows))
    - **Claim:** Several rows list "Timeout 300 s".
    - **Checks:** unit_consistency (time conversion) + repeated constant check

- **Verdict:** PASS
- **Notes:** Converted seconds to minutes and compared.

7. ✔ **C7** (Page 6, Section 4.2 (Fleet configuration and isolation))
   - **Claim:** Workstation specs include "two NVIDIA RTX PRO 6000 Blackwell GPUs (96 GB VRAM each)" and system RAM "512 GB".
   - **Checks:** parts_to_total (GPU aggregate VRAM)
   - **Verdict:** PASS
   - **Notes:** Computed total VRAM and checked optional inequality RAM > total VRAM.
8. ✔ **C8** (Page 8, Section 6.1 (Governing equation discovery))
   - **Claim:** Grid is "128^3" across "10 time slices".
   - **Checks:** cheap_recomputation (power and multiplication)
   - **Verdict:** PASS
   - **Notes:** Exact integer recomputation of 128^3 and multiplication by time slices.
9. ⚠ **C9** (Page 8, Section 6.1 (Governing equation discovery))
   - **Claim:** “constructed a library of 66 candidate terms”.
   - **Checks:** repeated_constant_cross_check (within-document references)
   - **Verdict:** UNCERTAIN
   - **Notes:** Within-document cross-occurrence check requires full document text; not available in PAYLOAD.
10. ✔ **C10** (Page 8, Section 6.1 (Governing equation discovery))
    - **Claim:** R^2 is reported as a range: "$R2 = 0.57$–0.71".
    - **Checks:** range_validity (bounded metric)
    - **Verdict:** PASS
    - **Notes:** Checked metric bounds and ordering.
11. ✔ **C11** (Page 8, Section 6.2 (Simulation mismatch detection))
    - **Claim:** "top 3 components capturing 97.35% of variance".
    - **Checks:** percentage_bounds
    - **Verdict:** PASS
    - **Notes:** Checked percentage bounds.
12. ✔ **C12** (Page 8, Section 6.2 (Simulation mismatch detection))
    - **Claim:** Partial AUC reported as "0.1488".
    - **Checks:** metric_bounds
    - **Verdict:** PASS

- **Notes:** Checked metric bounds $[0, 1]$.

13. ⚠ **C13** (Page 8, Section 6.2 (Simulation mismatch detection) vs Page 6 Table 1)
    - **Claim:** Section 6.2 says the paper was produced by "denario-3 (Claude Sonnet 4.6 with GPU access)"; Table 1 lists Scientist 3 with GPU "RTX PRO 6000" and model "Claude Sonnet 4.6".
    - **Checks:** cross_section_entity_consistency
    - **Verdict:** UNCERTAIN
    - **Notes:** Cross-section entity consistency requires extracted Table 1 text mapping scientist→(model,gpu); not available beyond isolated fields.
14. ✔ **C14** (Page 8, Section 6.3 (Multi-frequency analysis of tSZ maps))
    - **Claim:** Blind source detection yielded "200 candidates above 5σ"; Bullet Cluster recovered at "49σ".
    - **Checks:** unit-consistent comparison (sigma thresholds)
    - **Verdict:** PASS
    - **Notes:** Checked sigma inequality and that count is a non-negative integer.
15. ✔ **C15** (Page 8, Section 6.3 (Multi-frequency analysis of tSZ maps))
    - **Claim:** Spectral diagnostics across "90, 150, and 220 GHz" bands.
    - **Checks:** count_and_order check
    - **Verdict:** PASS
    - **Notes:** Checked count=3 and strict increasing order.
16. ✔ **C16** (Page 8, Section 6.3 (Multi-frequency analysis of tSZ maps))
    - **Claim:** “This 18-page paper with 16 figures …”
    - **Checks:** integer_relation (figures ≤ pages)
    - **Verdict:** PASS
    - **Notes:** Sanity inequality figures ≤ pages.
17. ✔ **C17** (Page 8, Section 7 (Monitoring and Cost Transparency))
    - **Claim:** Dashboard polls the supervised fleet "every 10 seconds".
    - **Checks:** unit_consistency (frequency conversion)
    - **Verdict:** PASS
    - **Notes:** Converted polling period to polls/min and compared.
18. ✔ **C18** (Page 9, Table 3 (Per-paper costs))
    - **Claim:** Table 3 lists three papers with total costs \$0.61, \$2.40, and \$4.10.
    - **Checks:** cheap_recomputation (sum, mean, min/max)
    - **Verdict:** PASS
    - **Notes:** Computed sum/mean/min/max; diff_* refers to sum vs 7.11. Mean checked within 0.01 as instructed.

19. ⚠ **C19** (Page 9, Table 3 (Per-paper costs) vs Page 8, Sections 6.1–6.3)
    - **Claim:** The three example papers in Sections 6.1–6.3 correspond to PX:2604.00016, PX:2604.00009, PX:2604.00015, which appear in Table 3.
    - **Checks:** cross_section_identifier_match
    - **Verdict:** UNCERTAIN
    - **Notes:** Requires Table 3 paper IDs to compare sets; PAYLOAD does not include Table 3 IDs (only costs).
20. ⚠ **C20** (Page 7, Table 2 + Page 6, Section 4.2)
    - **Claim:** Table 2 lists 3 connected systems; text states supervised denario-i fleet runs $N = 12$ scientists.
    - **Checks:** cross_table_consistency (counts stated in different locations)
    - **Verdict:** UNCERTAIN
    - **Notes:** Document-wide duplicate-number consistency requires full document text; not available in PAYLOAD.

Limitations

- Only parsed text provided from the PDF was used; no additional PDF structure (e.g., exact table cell boundaries) was available beyond the text transcript.
- No checks rely on external URLs, repositories, datasets, runtime logs, or executing the described systems; such claims are listed as unverified.
- No figure value extraction was performed; checks avoid reading plot pixels or inferring quantitative values from images.
- Some candidates (e.g., repeated-constant scans) are only actionable if you run text-search over the full PDF text; the provided transcript may omit formatting nuances.
- Some checks were uncertain because required quantities or structured table content were not available in the provided payload (e.g., DB size vs number of papers; Table 3 paper IDs for cross-matching; Table 1 scientist-to-(model,GPU) mapping for cross-section validation; full-document text for duplicate-number consistency).