CED-007 — Referee Review of Test Consistency

1. What This Review Is—and Is Not

This review evaluates the **consistency between two independent executions** of the Cosmological Explanation Diagnostic Audit (CEDA) applied to the same case study:
Janssen & Prokopec (2008), "Implications of the graviton one-loop effective action on the dynamics of the Universe."

It is **not** a re-audit of the paper, **not** a validation of CEDA's foundational premises, and **not** an assessment of the paper's scientific merit. It is a **reproducibility check**:

Does the protocol yield the same diagnostic classification when applied independently, using the same pre-declared cards and rules?

Accordingly, this review does **not** revisit whether the diagnostics are "fair" or whether the verdict is "correct." It evaluates alignment between runs in:

- final verdict classification,
- individual diagnostic outcomes,
- evidence citations (restricted to author statements), and
- regime scoping.

Any divergence would indicate protocol ambiguity or analyst discretion. Alignment supports auditability.

---

2. Why This Comparison Was Performed

This comparison was prompted by the availability of a second, independent CEDA execution ("Diagnostic Report") suitable for direct cross-check against an earlier run ("Methodology and Results").

The earlier run:

- was a structured application based strictly on the Model, Translation, and Diagnostic Cards,

- emphasized pre-diagnostic discipline to minimize bias,
- yielded an **Ambiguous** verdict with specific diagnostic flags.

The later report:

- appears as an independent execution in the same format,
- cites the same author-acknowledged limitations as evidence,
- invokes the same pre-declared failure criteria.

Together, these runs test CEDA's central claim: **reproducibility under fixed protocol, without interpretive drift.**

---

3. Pre-Review Discipline

Before comparison, all artifacts from both runs were aligned:

- **Model Card (both runs):**
  Identical capture of author aims, mechanisms, regimes, and limitations (e.g., constant-ε assumption, state sensitivity, lack of nonperturbative control).
- **Translation Card (both runs):**
  Identical mappings (e.g., "secular growth" → time-nonlocal contribution; no claim of scheme independence).
- **Diagnostic Card (Pre-Run, both runs):**
  Identical declarations:
  - D2 designated as the primary discriminator,
  - admissible variations limited to those discussed by the authors,
  - failure conditions triggered by scheme fragility or retuning.

No post-hoc adjustments were made.
Evidence in both runs was restricted to direct quotations or close paraphrases from the audited paper.

---

4. Diagnostics Compared (and Not Compared)

Both runs executed the same diagnostics:

- **D2 — Coarse-Graining Stability**
  Compared for outcome and supporting evidence.

- **D3 — Exchange-Term Provenance**
  Compared for conditional pass status and rationale.

- **C1 — Coupling Provenance & Redundancy**
  Compared for failure on functional freedom / degeneracy.

- **S1 — Scheme / State Dependence Classification**
  Compared for assignment to S1-D (scheme/state fragile).

The following were **not** applied in either run:

- **D1** (no horizon-only claim),

- **D4** (disabled due to D2 failure).

No divergence occurred in diagnostic selection or sequencing.

---

5. How to Read the Comparison

The verdicts **match**:

Ambiguous (scheme/state fragile; regime-limited)

This agreement reflects the following shared findings:

- The one-loop effect is credited as formally derived within the declared constant-ε regime.

- Broader claims (Λ screening, late-time attractors) are classified as unstable due to author-acknowledged issues:

    - unresolved resummation,

- counterterm / scheme sensitivity,
- vacuum and state-evolution assumptions,
- extrapolation beyond controlled ε.

Evidence alignment:
Both runs cite the same author warnings (e.g., reliability breakdown near poles, need for resummation, threatened vacuum assumptions).

Differences:
Minor phrasing and emphasis (e.g., one run foregrounds “predictive compression not paid,” the other “functional freedom / degeneracy risk”).
No material difference appears in diagnostic flags or verdict logic.

---

6. What Would Count as a Reproducibility Failure

For clarity, reproducibility would be undermined if:

- diagnostic outcomes differed (e.g., D2 passing in one run and failing in another) without different evidence,
- undeclared variations were introduced in one run (e.g., ignoring author-flagged state dependence),
- verdicts diverged due to implicit bias rather than protocol differences.

None of these occurred.
Alignment holds across all diagnostic layers.

---

7. Scope of Further Evaluation

This review does **not** assess:

- whether CEDA over-weights author self-critique,
- whether the paper’s limitations are resolvable,

- whether related models could pass under extended control.

It assesses only:

- consistency between independent executions,
- traceability of evidence to pre-declared criteria,
- whether admissible alternative translations would flip outcomes.

No such flips are anticipated; both runs are tightly bound to the authors' own regime declarations.

Closing Note

This comparison affirms **CEDA's auditability**: independent applications yield consistent diagnostic outputs when constrained by the same protocol and evidence base.

It is a transparency check, not a defense of the framework or the paper.
Should broader testing reveal systematic divergence, or should the protocol prove too rigid or permissive, that feedback would motivate revision.

CEDA's claim here is limited but concrete:
when rules are fixed, results converge.