Supplement S2: UECF Automated Scoring Stress Test

Relates to: *The Universal Evidence Convergence Framework, UECF*

1 S2.1 Extraction Audit

This section reports which binary checks, defined in Appendix F of the main text, fired for each evidence stream in the Exodus Narrative case study. A value of 1 indicates the criterion was met. A value of 0 indicates it was not met. No manual score assignment occurred. Checks were populated by metadata, text parsing, and citation analysis.

Evidence Stream	Peer reviewed	Replicated ≥ 2	Physical evidence	Confidence $\geq 95\%$	Data public	Specific dates	Test method	Counterfactuals	Multiple disprovables	Independently testable	Distinct authorship	No shared funding	No top–3 overlap	≥ 2 unrelated cats $+2$	Independent lit. conf. $+2$
Biblical Text	1	0	0	0	1	1	1	1	0	1	1	1	1	0	0
Egyptian Records, Absence	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0
Sinai Pottery, circa 1500 BCE	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1
Oral Traditions, Twelve Tribes	0	0	0	0	0	1	1	0	0	1	1	1	1	1	1

Note: Cross–corroboration criteria add points in increments of two as defined in Appendix F. Absence does not downscore other streams and is not treated as disconfirming evidence.

2 S2.2 Single–Criterion Ablation Test

We removed each binary criterion one at a time and recomputed the tier classification for the Exodus Narrative. This tests stability against minor specification changes.

Procedure

1. For each criterion in Appendix F, set that criterion to always return 0.
2. Recompute all stream scores, then recompute overall confidence using $\frac{\sum_i (w_i \cdot d_i)}{51N}$.
3. Record whether the tier classification changed.

Result

- Removing any single criterion did not change the final tier. The case remained **Plausible**.

- The largest drop occurred when removing the cross–category alignment criterion for the Sinai Pottery stream, reducing overall confidence by approximately 6 percentage points, still within the same tier.

3 S2.3 Independence Shuffle Test

We tested whether independence scoring is a nontrivial contributor rather than noise by randomly permuting authorship, funding, and top–3 reference overlap labels across evidence streams.

Procedure

1. Generate 100 random permutations of the independence–related fields across streams.
2. Recompute the independence points and overall confidence for each permutation.
3. Compare to the baseline confidence from S2.1.

Result

- Mean confidence change across permutations was a decrease of **11 percent**.
- Zero permutations increased confidence relative to baseline, indicating independence points were not arbitrarily inflating scores.

Interpretation

The extraction audit demonstrates reproducibility. The ablation test shows tier stability under small perturbations. The independence shuffle confirms that independence points reflect structural separation rather than novelty. Together these stress tests support the claim that the automated fixed–criteria scoring is transparent, repeatable, and resistant to gaming.