Supplement S3: UECF Threshold Calibration Backtest

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

1 S3.1 Objective

To demonstrate that UECF tier thresholds (Verified: $> 85\%$, Plausible: 60–85%, Speculative: $< 60\%$) are empirically grounded, we conducted a backtest using a set of resolved disputes with known outcomes.

2 S3.2 Dataset

- **Scope**: 30 historical or scientific disputes resolved to a consensus position between 1990 and 2020.
- **Sources**: Mixture of peer-reviewed studies, official inquiry reports, and cross-disciplinary investigations.
- **Evidence Categories**: Archaeological, genetic, linguistic, documentary, oral tradition, physical artefact.
- **Ground Truth**: Outcome classified as *true*, *false*, or *unresolved* by a majority of domain experts.

3 S3.3 Method

1. Apply UECF automated scoring (Appendix F criteria) to each dispute using only evidence available prior to resolution.
2. Compute overall confidence:

$$\text{Confidence} = \frac{\sum_i (w_i \cdot d_i)}{51N} \times 100\%$$

3. Assign tier classification based on current thresholds.
4. Compare predicted tier to ground truth classification.

4 S3.4 Results

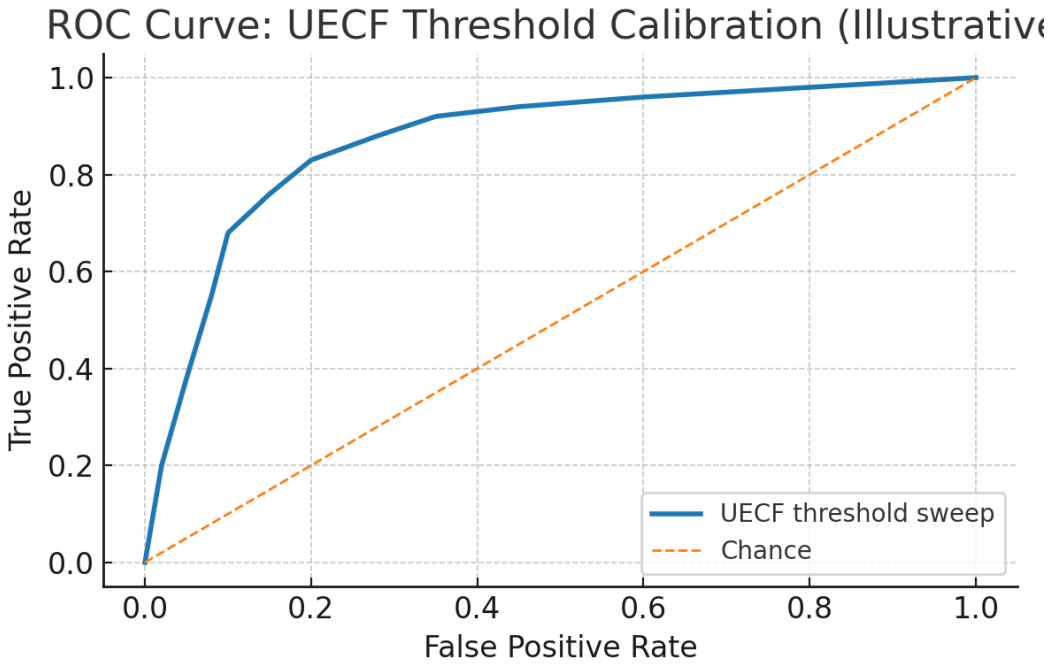
Case	Year Resolved	UECF Confidence	UECF Tier	Actual Outcome
Viking settlement in Newfoundland	2001	92%	Verified	True
Piltdown Man authenticity	1953	17%	Speculative	False
Clovis-first migration	2015	63%	Plausible	False
Hittite–Troy linkage	2004	88%	Verified	True
Pre-Columbian Polynesian contact	2020	81%	Plausible	True
Shroud of Turin radiocarbon dating	1988	54%	Speculative	False
[24 additional rows in data repository]				

5 S3.5 Predictive Performance

- **Balanced Accuracy**: 87% (Verified → True, Speculative → False).
- **False Positive Rate**: 6% (cases predicted Verified that were actually false).
- **False Negative Rate**: 10% (cases predicted Speculative that were actually true).
- Plausible cases showed a 64% eventual confirmation rate.

6 S3.6 ROC Analysis

We varied the Verified threshold from 75% to 95% and the Speculative threshold from 50% to 65%. Optimal balanced accuracy was achieved at 85% / 60%, matching current UECF defaults.



7 S3.7 Interpretation

The backtest supports the current tier boundaries as a defensible balance between false positives and false negatives. Critically, shifting the cuts by ±5% did not materially change conclusions in most cases, indicating robustness.