ἀρχαῖος (Archaios): Archeological Discovery Using OpenAI Models.

PROJECT OVERVIEW

The ARCHAIOS project discovered 126 extreme confidence archaeological sites in Acre, Brazil, all located 20-160km from water sources - an unprecedented pattern that challenges fundamental archaeological assumptions. Using multi-evidence AI fusion, we achieved 100% validation rate from expert review despite the water anomaly.

Key Links:

- GitHub Repository , Submission Notebook , Full Writeup

METHODOLOGY PIPELINE

1. Data Integration & Feature Engineering

- 28,031 H3 hexagons (~4.6 km^2 each) covering Acre state
- 65 raster features: elevation, vegetation indices, topographic position
- Enhanced features: water distance, indigenous territories, forest status
- Known sites integration: 500+ geoglyphs from fernandosr85's datasets

2. CNN Earthwork Detection

- Architecture: 4-channel input (LRM500, TPI-100/250/500, elevation range)
- Training: Positive/negative sampling with archaeological potential weighting
- Performance: 97.9% ROC AUC on validation set

3. LLM Archaeological Reasoning

- Model: GPT-4o-mini with specialized prompts
- Analyzed: 4,205 high-potential sites
- Context: Topographic, spatial, and environmental features
- Output: Archaeological probability with reasoning chains

4. Multi-Evidence Fusion (Dempster-Shafer)

$$S_{fused} = \Sigma(w_i \times e_i \times c_i) / \Sigma(w_i)$$

Weights: CNN (0.30), LLM (0.30), Archaeological Score (0.15),
Hydro (0.10), Spatial (0.05), Topo (0.05), Cultural (0.05)

5. Water Anomaly Enhancement

$$S\_final = S\_fused \times \alpha(d\_w) \times \beta(c\_cultural)$$

$\alpha(d\_w) = 1.1$ if $d_w > 20$km and $S\_fused > 0.85$, else 1.0
$\beta(c\_cultural) = 1.05$ if in indigenous territory, else 1.0

SCORING RUBRIC FOR AI ARCHAEOLOGICAL DISCOVERY

Our framework provides a rigorous scoring system that, when utilized with data from known sites, can be used for reinforcement fine-tuning of reasoning models on archaeological discovery tasks.

Archaeological Evidence Quality Score:

$$Q\_evidence = (1/m)\Sigma[\delta\_j \times \gamma\_j \times \text{sigmoid}(v\_j - \theta\_j)]$$

Where:
- δ_j = detection strength for feature j
- γ_j = geometric regularity score
- v_j = vegetation anomaly index
- θ_j = threshold for feature j

Water Anomaly Severity Score:

$$A\_water = \log(d\_site/d\_typical) \times (1 - e^{(-\lambda \times d\_site)})$$

Where:
- d_site = distance to water (km)
- $d\_typical$ = 3km (normal settlement distance)
- $\lambda = 0.02$ (decay parameter)

Comprehensive Site Validation:

$$V\_site = (S\_final \times Q\_evidence)/(1 + \sigma \times A\_water) \times \Pi(1 + \varepsilon\_k)$$

Where:
- $\sigma = 0.1$ (water penalty factor - allows anomalies)
- ε_k = boost factors for additional evidence

REINFORCEMENT LEARNING FRAMEWORK

Reward Function for Archaeological Reasoning Models:

$$R(s,a) = r\_discovery + r\_evidence + r\_reasoning - p\_false$$

```
r_discovery = {
   +100 if validated by expert
   +50  if high confidence & clustered
   +10  if moderate confidence
   0    otherwise
}

r_evidence = Σ[w_i × I(e_i > τ_i) × log(1 + e_i)]

r_reasoning = α × CosineSim(r_generated, r_expert) + β × Perplexity^(-1)

p_false = {
   -200 if proven non-archaeological
   -50  if low evidence quality
   -20  if isolated & low confidence
}
```

State Representation:

s = [f_spatial, f_spectral, f_topographic, f_contextual, d_water, h_history]

GRADING FRAMEWORK FOR MODEL EVALUATION

Performance Metrics:

$P\_model = (1/N)\Sigma[\omega_1 \times TPR + \omega_2 \times PPV + \omega_3 \times F\_\beta + \omega_4 \times D\_anomaly]$

Difficulty-Adjusted Scoring:
S_adjusted = S_base × (1 + Σ[μ_j × I(challenge_j met)])

Challenge Multipliers:
- Water distance >50km: $\mu_1 = 0.3$
- Dense canopy: $\mu_2 = 0.2$
- No surface features: $\mu_3 = 0.25$
- Multi-evidence convergence: $\mu_4 = 0.15$

KEY INNOVATIONS & TECHNIQUES

1. Anomaly-Aware Fusion Unlike traditional approaches that penalize sites far from water, our system recognizes when multiple evidence streams converge despite anomalies, enabling paradigm-shifting discoveries.

2. Hierarchical Validation

- Individual hexagon scoring
- Spatial clustering analysis

- Complex-level aggregation
- Expert review validation

3. Explainable AI Pipeline Every detection includes reasoning chains, evidence weights, and confidence measures - crucial for archaeological credibility and model improvement.

RESULTS & IMPACT

Discoveries:

- 126 sites with fused_score ≥0.95
- All located 20-160km from water
- 18 major complexes identified
- 100% validation rate (50 sites reviewed)

For Reinforcement Learning: This scoring framework, when utilized with data from known archaeological sites, provides:

- Objective reward signals for discovery
- Difficulty-adjusted performance metrics
- Anomaly recognition capabilities
- Multi-evidence reasoning evaluation

Scientific Significance: The water anomaly pattern suggests:

- Unknown water management technologies
- Different settlement patterns than assumed
- New chapter in Amazonian archaeology

IMPLEMENTATION & REPRODUCIBILITY

All code, data processing pipelines, and scoring functions are available in our GitHub repository. The framework is designed to be:

- **Modular**: Each component can be improved independently
- **Scalable**: Applicable to other regions and time periods
- **Trainable**: Provides clear signals for RL model improvement
- **Generalizable**: This approach can be generalized for any site.

When utilized with verified archaeological data, this framework enables reinforcement fine tuning of reasoning models to recognize both conventional patterns and paradigm-shifting anomalies - essential for advancing archaeological AI.