# AI Audit Tool Evaluation Research Methodology

## Project Overview

### Objective

To conduct an independent, unbiased evaluation of AI-powered smart contract audit tools across multiple categories and contract complexities. This research aims to provide the community with transparent data on tool effectiveness rather than commercial recommendations.

### Research Questions

- How accurate are current AI audit tools in identifying real vulnerabilities?
- What is the precision vs recall trade-off across different tools?
- Do tool performance vary by contract category or complexity?
- What types of vulnerabilities do AI tools commonly miss or misidentify?

## Methodology

### Dataset Composition

**Sample Size**: 9 contracts (pilot study)

- 3 High complexity (1500+ SLOC)
- 3 Medium complexity (800-1500 SLOC)
- 3 Low complexity (<800 SLOC)

**Contract Categories**:

- Dexes
- Lending protocols
- Yield farming
- Liquid staking
- Stablecoins
- Real World Assets (RWA)

**Baseline Truth**: Public audit reports for each contract, assuming reported vulnerabilities represent the complete set of issues.

### AI Tools Under Evaluation

[Tools will be disclosed in final results with full transparency]

### Evaluation Framework

**Primary Metrics**:

- **Precision**: Valid findings / Total findings (measures spam/noise)
- **Recall**: Found vulnerabilities / Known vulnerabilities (measures coverage)

**Scoring System** (per finding, 0-4 scale):

- **4**: Valid vulnerability, accurate description, correct severity
- **3**: Valid vulnerability, mostly accurate description, minor issues
- **2**: Valid vulnerability with poor description OR reasonable false positive
- **1**: Clear false positive but shows code understanding
- **0**: Nonsense/irrelevant finding

### Evaluator Selection

**Qualification Requirements**:

- Minimum 2-3 public contest judging experiences
- Demonstrated smart contract security expertise
- No employment or consulting relationships with evaluated AI tool companies
- No involvement in creating/auditing the contracts being evaluated

**Conflict of Interest Policy**:

- Evaluators cannot work for any AI audit tool company under evaluation
- Evaluators cannot have authored or previously audited the test contracts
- Full disclosure of any potential conflicts before participation

**Compensation**: $30-50 per contract evaluation (research/academic rates)

## Process Design

### Phase 1: Recruitment (2 days)

1. Post recruitment call in security researcher communities
2. Screen candidates against qualification criteria
3. Confirm availability and sign conflict of interest agreements
4. Provide evaluation guide and scoring rubric

### Phase 2: Tool Execution

1. Run all AI tools against the 9-contract dataset
2. Collect and standardize all findings
3. Anonymize findings (remove tool identifiers)
4. Randomly distribute findings to evaluators

### Phase 3: Evaluation (1 week)

1. Evaluators independently score findings using 0-4 scale
2. Blind evaluation - evaluators don't know which tool generated which finding
3. Each finding evaluated by multiple evaluators for reliability
4. Daily check-ins to ensure progress and answer questions

### Phase 4: Quality Control

**Disagreement Resolution**:

- Initial disagreements handled through group discussion
- Persistent disagreements resolved by majority vote among all evaluators
- Final arbitration by research coordinator when needed

**Inter-rater Reliability**:

- Track agreement rates between evaluators
- Calculate Cohen's kappa for scoring consistency
- Flag findings with high disagreement for additional review

### Phase 5: Analysis & Publication

**Data Analysis**:

- Calculate precision/recall for each AI tool
- Break down performance by contract category and complexity
- Identify common failure patterns and tool strengths
- Statistical significance testing where applicable

**Publication Plan**:

1. **Initial Post**: High-level results and key findings
2. **Blog Series**: Deep-dive analysis by category and tool performance
3. **Raw Data**: Anonymized dataset and scoring results for community review

**Transparency Commitments**:

- Full methodology disclosure
- Complete tool identification (no anonymization in results)
- Evaluator backgrounds and selection process
- Sponsor acknowledgment and funding details
- Raw data availability for independent verification

## Limitations & Scope

### Known Limitations

- Small sample size (pilot study)
- Baseline assumes public audits found all vulnerabilities
- Point-in-time evaluation (tool capabilities evolve)
- Limited to specific contract categories

### Future Expansion

Based on pilot results, potential expansion to:

- Larger contract dataset (50+ contracts)
- Additional categories and complexity ranges
- Longitudinal studies tracking tool improvement
- Cross-chain protocol evaluation

## Ethical Considerations

### Research Independence

- No commercial relationships with evaluated tools

- Academic/research compensation rates
- Open publication of all results
- Community benefit focus over commercial gain

### Fair Evaluation

- Standardized evaluation criteria
- Multiple evaluator perspectives
- Transparent disagreement resolution
- Tool-agnostic methodology

## Expected Outcomes

### Deliverables

1. Comprehensive performance metrics for each AI audit tool
2. Category and complexity-based performance analysis
3. Identified strengths and weaknesses of current AI approaches
4. Recommendations for tool improvement and user guidance
5. Open dataset for community research

### Community Impact

- Evidence-based tool selection guidance
- Identification of current AI audit limitations
- Baseline for tracking industry progress
- Framework for future comparative studies

---

**Sponsor Acknowledgment**: This research is made possible by [Sponsor Name], with full editorial independence maintained throughout the evaluation process.

**Timeline**:

- Recruitment: 2 days
- Evaluation: 1 week
- Analysis & Publication: 2-3 weeks

**Contact**: [Research coordinator contact information]