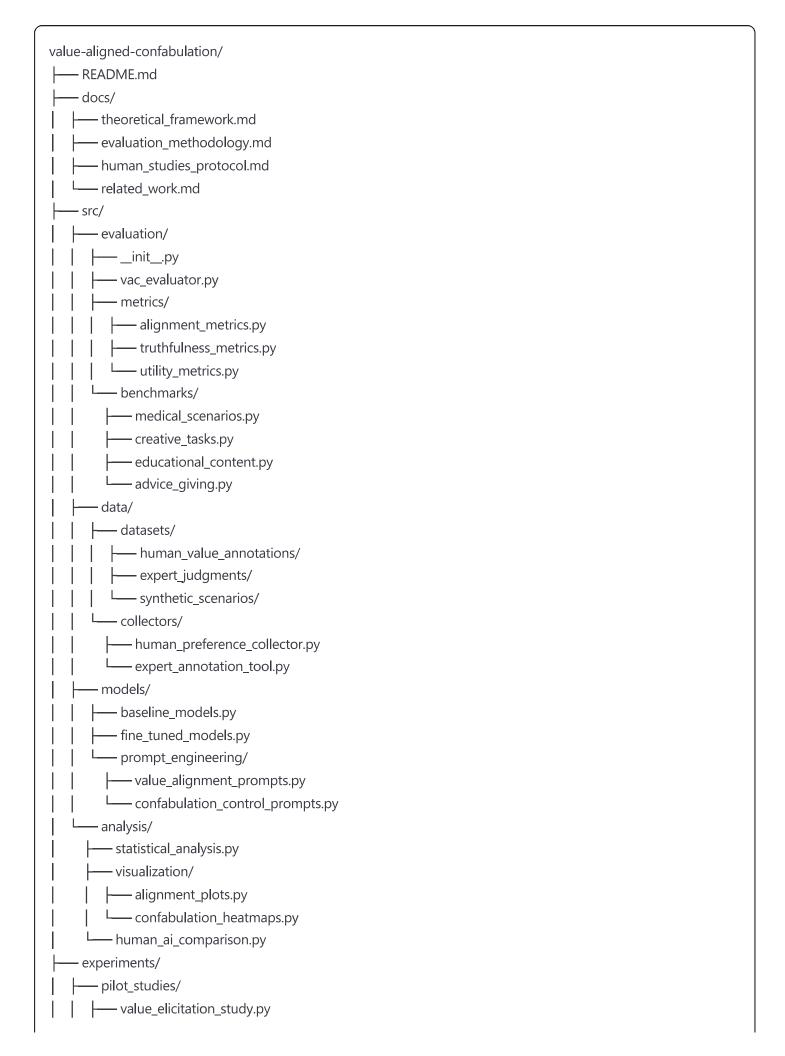
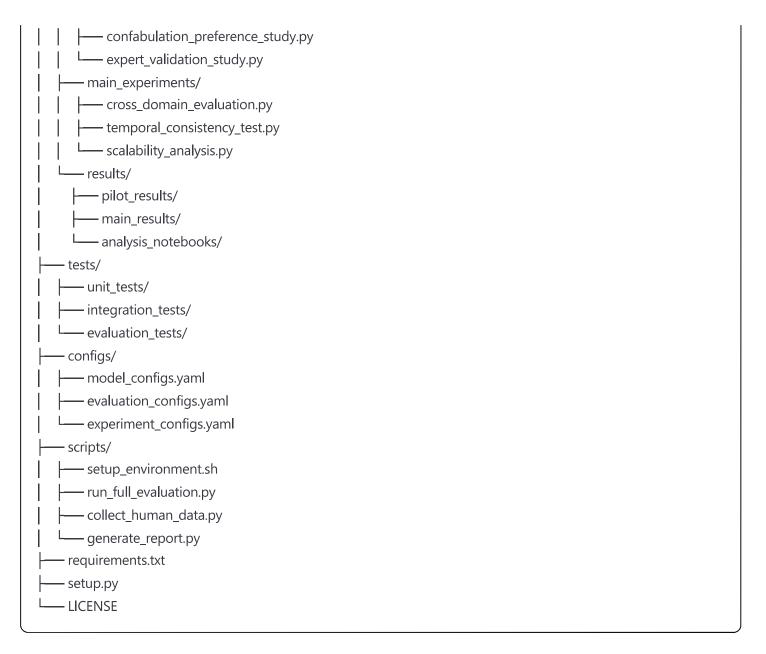
Value-Aligned Confabulation (VAC) Research Repository

Repository Structure





Core Components

1. Theoretical Framework (docs/theoretical_framework.md)

Key Concepts:

- Value-Aligned Confabulation (VAC): LLM outputs that are factually ungrounded but align with human values and serve beneficial purposes
- Harmful Hallucination: Factually incorrect outputs that mislead or cause harm
- Truthfulness-Utility Trade-off: The balance between factual accuracy and beneficial outcomes

Research Questions:

- 1. Can LLMs learn to confabulate in ways that align with human values?
- 2. How do we measure the alignment between beneficial confabulation and truthfulness?

3. What contextual factors determine when confabulation becomes harmful vs. helpful?

2. Evaluation Methodology (docs/evaluation_methodology.md)

Multi-Dimensional Assessment:

- Alignment Score: How well confabulation aligns with human values
- Truthfulness Score: Factual accuracy of claims
- Utility Score: Practical benefit of the output
- **Transparency Score**: How well the model expresses uncertainty

Evaluation Domains:

- Medical Advice: High stakes, low tolerance for harmful confabulation
- **Creative Writing**: High tolerance, high value for beneficial confabulation
- Educational Content: Balanced approach, pedagogically useful speculation
- Personal Advice: Context-dependent, requires value alignment

3. Core Evaluation Framework (src/evaluation/vac_evaluator.py)

Main Evaluator Class:

```
class ValueAlignedConfabulationEvaluator:

def __init__(self, config):

self.alignment_metrics = AlignmentMetrics()

self.truthfulness_metrics = TruthfulnessMetrics()

self.utility_metrics = UtilityMetrics()

def evaluate_response(self, prompt, response, context):

# Multi-dimensional evaluation

pass

def compute_vac_score(self, alignment, truthfulness, utility):

# Weighted combination based on context

pass
```

4. Benchmark Scenarios (src/evaluation/benchmarks/)

Medical Scenarios:

• Patient questions where speculation could be harmful

- Educational medical content where reasonable inference is helpful
- Emergency situations requiring clear uncertainty communication

Creative Tasks:

- Story completion where confabulation enhances narrative
- Brainstorming sessions where speculation drives innovation
- Artistic interpretation where subjective "truth" is valuable

Educational Content:

- Explaining complex concepts through analogies
- Filling knowledge gaps with pedagogically useful speculation
- Historical scenarios requiring reasonable inference

5. Human Studies Protocol (docs/human_studies_protocol.md)

Value Elicitation Study:

- Collect human judgments on when confabulation is acceptable
- Map contextual factors that influence preference
- Establish baseline human values for alignment

Preference Collection:

- Pairwise comparisons between different types of confabulation
- Context-dependent preference modeling
- Expert vs. lay person preference differences

6. Key Experiments (experiments/)

Pilot Studies:

- Value elicitation from diverse human populations
- Initial model evaluation on core scenarios
- Expert validation of evaluation framework

Main Experiments:

- Cross-domain evaluation of VAC across different contexts
- Temporal consistency testing for long conversations

Scalability analysis for different model sizes

7. Analysis Tools (src/analysis/)

Statistical Analysis:

- Correlation between alignment and utility scores
- Domain-specific performance patterns
- Human-Al agreement analysis

Visualization:

- Alignment-truthfulness scatter plots
- Domain-specific performance heatmaps
- Temporal consistency tracking

Initial Implementation Priority

Phase 1: Foundation (Weeks 1-2)

- 1. Set up repository structure
- 2. Implement core evaluation framework
- 3. Create initial benchmark scenarios
- 4. Develop basic metrics

Phase 2: Human Studies (Weeks 3-4)

- 1. Design and run value elicitation study
- 2. Collect expert judgments
- 3. Establish baseline human preferences
- 4. Validate evaluation methodology

Phase 3: Model Evaluation (Weeks 5-6)

- 1. Evaluate baseline models (GPT-4, Claude, etc.)
- 2. Test across different domains
- 3. Analyze alignment-truthfulness trade-offs
- 4. Generate initial results

Phase 4: Analysis & Iteration (Weeks 7-8)

- 1. Statistical analysis of results
- 2. Refine evaluation metrics based on findings
- 3. Prepare research publication
- 4. Plan next phase of research

Key Research Deliverables

- 1. **VAC Evaluation Framework**: Standardized methodology for assessing value-aligned confabulation
- 2. **Benchmark Dataset**: Curated scenarios for testing VAC across domains
- 3. Human Values Database: Annotated preferences for confabulation in different contexts
- 4. **Model Performance Analysis**: Comprehensive evaluation of current LLMs on VAC tasks
- 5. **Research Paper**: "Value-Aligned Confabulation: Moving Beyond Binary Truthfulness in LLM Evaluation"

Collaboration Opportunities

- Academic Partnerships: Al safety researchers, cognitive scientists, philosophers
- Industry Engagement: Model developers interested in alignment research
- **Community Involvement**: Open-source contributors, evaluation researchers
- Policy Implications: Al governance researchers, ethicists

This structure provides a comprehensive foundation for rigorous research into value-aligned confabulation while remaining practical and implementable.