

A Novel Framework for Semantic-Based LLM Reasoning Assessment

Tyler Williams, VANTA Research 2025

Abstract

Standard Large Language Model (LLM) evaluation frameworks often fail to capture true reasoning capabilities due to their reliance on exact format matching and token probability scoring. This paper introduces the VANTA Research Reasoning Evaluation (VRRE), a semantic understanding framework that measures reasoning ability through intelligent response parsing and partial credit scoring systems.

Our validation demonstrates that VRRE successfully identifies reasoning improvements that standard benchmarks miss—specifically, a 2.5x improvement in logical reasoning performance between model variants that showed identical scores (22% accuracy) on traditional benchmarks.

Introduction

The Problem with Current Evaluation

Traditional LLM benchmarks face several critical limitations:

1. **Format Dependency:** Requiring exact answer formats ("Yes"/"No") rather than understanding semantic meaning
2. **Binary Scoring:** Missing nuanced reasoning quality in responses
3. **Probability Bias:** Using token probabilities instead of semantic correctness
4. **Limited Reasoning Coverage:** Focusing on narrow task types rather than comprehensive reasoning assessment

The VRRE Solution

VRRE addresses these limitations through:

- **Semantic Answer Extraction:** Intelligent parsing of natural language responses
- **Partial Credit System:** Rewarding reasoning process even with incorrect final answers
- **Multi-Domain Assessment:** Covering boolean logic, mathematics, reading comprehension, and formal reasoning
- **Confidence Scoring:** Providing reliability metrics for extracted answers

Methodology

Semantic Answer Extraction Algorithm

```
def extract_boolean_answer(self, response: str) -> Tuple[str, float]:
    """Extract yes/no answers with confidence scoring"""
    yes_patterns = [
        r'\byes\b', r'\btrue\b', r'\bcorrect\b', r'\bvalid\b',
        r'we can conclude', r'it is valid', r'this is true'
    ]
    no_patterns = [
        r'\bno\b', r'\bfalse\b', r'\bincorrect\b', r'\binvalid\b',
        r'cannot conclude', r'not valid', r'fallacy'
    ]

    # Pattern matching with reasoning indicator boost
    reasoning_boost = 0.1 * sum(1 for indicator in reasoning_indicators
                                if indicator in response.lower())

    # Confidence calculation based on pattern strength and reasoning quality
    confidence = min(0.95, 0.5 + pattern_differential * 0.15 +
                    reasoning_boost)
```

Scoring System

Full Credit (1.0): Correct answer with sound reasoning

Partial Credit (0.2-0.3): Incorrect answer but demonstrates reasoning process

No Credit (0.0): Wrong answer with poor/no reasoning

Task Categories

1. **Boolean Logic:** Syllogisms, logical fallacies, deductive reasoning
2. **Mathematical Reasoning:** Arithmetic, geometry, word problems
3. **Reading Comprehension:** Passage-based inference tasks
4. **Formal Logic:** Validity assessment, premise evaluation

Comprehensive Multi-Model Validation Study

Experimental Design

We conducted extensive validation across five diverse models representing different architectures, sizes, and training paradigms:

- **phi3.5** (2.2GB): Microsoft's latest efficient reasoning model
- **qwen2:0.5b** (352MB): Ultra-lightweight Chinese architecture

- **mistral:7b** (4.4GB): Established baseline model
- **apollo-reasoning-enhanced** (4.4GB): Enhanced with logical reasoning guidelines
- **apollo-system-prompt** (4.4GB): System prompt approach for identity preservation

VRRE vs Standard Benchmark Performance

Overall Reasoning Quality Rankings

Rank	Model	VRRE Score	Expected Standard*	Efficiency (Quality/GB)
1	phi3.5	59.7%	MMLU: 69%, GSM8K: 87%	27.1
2	mistral:7b	53.6%	MMLU: 60%, GSM8K: 52%	12.2
3	qwen2:0.5b	53.6%	MMLU: 49%, GSM8K: 58%	152.3
4	apollo-system	47.5%	Identity: ~95%, Conv: ~90%	10.8
5	apollo-enhanced	46.6%	Identity: ~95%, Conv: ~90%	10.6

*Approximate expected performance on standard benchmarks

Novel Insights Invisible to Standard Benchmarks

1. The Size Efficiency Revolution

Discovery: Qwen2:0.5b achieves remarkable reasoning efficiency at 352MB

- **VRRE Efficiency:** 152.3 quality points per GB (12.5x more efficient than Mistral)
- **Standard Benchmark Blind Spot:** Raw accuracy scores would show Mistral significantly ahead
- **Research Implication:** Architectural efficiency matters more than parameter count for reasoning tasks

2. Domain-Specific Reasoning Imbalances

Discovery: Severe math vs logic performance disparities across models

Model	Math Reasoning	Logic Reasoning	Domain Gap
phi3.5	73.3%	23.3%	+50.0%
mistral:7b	46.7%	23.3%	+23.4%
qwen2:0.5b	46.7%	26.7%	+20.0%
apollo-enhanced	20.0%	26.7%	-6.7%
apollo-system	20.0%	26.7%	-6.7%

Standard Benchmark Limitation: Domain-specific tests (GSM8K, ARC) don't reveal cross-domain consistency

VRRE Insight: Models exhibit severe reasoning domain imbalances invisible to specialized benchmarks

3. Training Optimization Mismatch

- Discovery:** Apollo variants significantly underperform despite sophisticated reasoning training
- **VRRE Reality:** Apollo-enhanced achieves only 23.8% accuracy
 - **Expected Performance:** 65%+ on conversational/identity tasks
 - **Training Insight:** Optimization for conversational consistency ≠ logical reasoning capability
 - **Research Implication:** Need for reasoning-specific evaluation during training

4. Speed-Quality Trade-off Revelation

Discovery: Practical deployment efficiency varies dramatically

Model	Quality/Second	Optimal Use Case
qwen2:0.5b	44.7	Edge deployment reasoning
phi3.5	28.4	Math-intensive applications
mistral:7b	19.1	Balanced reasoning tasks
apollo-system	18.3	Conversational AI
apollo-enhanced	16.1	Identity-consistent tasks

Standard Benchmark Gap: Pure accuracy metrics ignore real-world inference constraints

VRRE Contribution: Actionable deployment guidance based on reasoning efficiency

Original Apollo Enhancement Validation

The Breakthrough Discovery

Initial Study: Two Apollo variants that showed identical standard benchmark scores

Benchmark	apollo-system-prompt	apollo-reasoning-enhanced	Difference
BoolQ	22%	22%	0%
PIQA	56%	56%	0%
ARC Easy	18-28%	18-28%	0%

VRRE Reveals Hidden Improvement

Category	apollo-system-prompt	apollo-reasoning-enhanced	Improvement
Overall	22.2% accuracy	55.6% accuracy	+2.5x
Boolean Logic	0%	50%	+∞
Mathematical	100%	100%	Equal
Reading Comp	0%	100%	+∞

Critical Finding: VRRE detected a 2.5x reasoning improvement invisible to standard benchmarks

The Logical Fallacy Test Case

Task: "All roses are flowers. Some flowers are red. Can we conclude that some roses are red?"

apollo-system-prompt: "yes" (incorrect - commits logical fallacy)

apollo-reasoning-enhanced: "unclear" (correct - avoids fallacy)

This demonstrates VRRE's ability to detect logical reasoning improvements that standard benchmarks miss entirely.

Comprehensive Results Analysis

Critical Gaps in Standard Evaluation Revealed

1. The Efficiency Blind Spot

Standard benchmarks evaluate raw accuracy but ignore computational efficiency for reasoning tasks:

- **VRRE Discovery:** Qwen2:0.5b achieves 53.6% reasoning quality at 352MB
- **Standard Expectation:** Mistral:7b (4.4GB) should significantly outperform

- **Reality:** Identical reasoning scores with 12.5x size difference
- **Implication:** Current benchmarks fail to identify efficient reasoning architectures

2. Domain Balance Invisibility

Existing benchmarks test domains in isolation, missing cross-domain reasoning consistency:

- **VRRE Discovery:** Phi-3.5 shows 50% gap between math (73.3%) and logic (23.3%)
- **Standard Approach:** GSM8K (87%) suggests strong reasoning, ARC misses logic weakness
- **Reality:** Severe domain imbalances in reasoning capability
- **Implication:** Need for integrated reasoning assessment across domains

3. Training Objective Mismatch Detection

Standard benchmarks can't detect when training optimizes for the wrong reasoning patterns:

- **VRRE Discovery:** Apollo variants score 23.8% despite sophisticated training
- **Standard Expectation:** Should score 65%+ on identity/conversation tasks
- **Reality:** Conversational training \neq logical reasoning capability
- **Implication:** Training evaluation must include reasoning-specific metrics

Why VRRE Succeeds Where Others Fail

Semantic vs Syntactic Understanding

```
# Standard Benchmark Approach
if response.strip().lower() == "no":
    score = 1.0
else:
    score = 0.0

# VRRE Approach
extracted_answer, confidence = semantic_extractor.extract_answer(response,
task_type)
score = calculate_partial_credit(extracted_answer, correct_answer,
reasoning_quality)
```

Partial Credit Recognition

- **Standard:** Binary right/wrong scoring
- **VRRE:** Recognizes reasoning process quality
- **Impact:** Captures incremental improvements in logical thinking

Cross-Architecture Discrimination

VRRE successfully ranked 5 diverse models with 18.7% accuracy spread, demonstrating:

- **Reliability:** 100% extraction success across all architectures
- **Sensitivity:** Clear performance tiers emerged
- **Robustness:** Consistent domain pattern recognition

Validation of Core Framework Claims

Claim 1: "Standard benchmarks miss reasoning improvements"

Validated: Apollo 2.5x improvement invisible to BoolQ, PIQA, ARC

Claim 2: "Semantic parsing outperforms format compliance"

Validated: 100% extraction success across 5 diverse architectures

Claim 3: "VRRE reveals practical deployment insights"

Validated: Qwen2:0.5b efficiency discovery changes deployment recommendations

Claim 4: "Cross-domain reasoning assessment needed"

Validated: Phi-3.5 domain imbalance invisible to specialized benchmarks

Implications for AI Research and Industry

For Model Development and Selection

Size-Efficiency Optimization

- **Discovery:** Qwen2:0.5b achieves competitive reasoning at 352MB
- **Implication:** Focus on architectural efficiency over parameter scaling
- **Application:** Edge deployment with meaningful reasoning capabilities

Domain-Balanced Training

- **Discovery:** Models show severe reasoning domain imbalances (50% gaps)
- **Implication:** Training must ensure cross-domain reasoning consistency
- **Application:** Multi-domain reasoning benchmarks during training

Training Objective Alignment

- **Discovery:** Conversational training ≠ logical reasoning capability
- **Implication:** Reasoning-specific evaluation needed during development
- **Application:** VRRE integration into training pipelines

For Evaluation Methodology Revolution

Beyond Format Compliance

Standard benchmarks must evolve from "exact format matching" to "semantic understanding":

```
# Current Standard Approach
def evaluate_response(response, expected_answer):
    return 1.0 if response.strip() == expected_answer else 0.0

# VRRE Semantic Approach
def evaluate_response(response, task):
    answer, confidence = extract_semantic_meaning(response, task.type)
    return calculate_reasoning_quality(answer, task.correct, response)
```

Process vs Product Assessment

- **Current:** Binary right/wrong scoring
- **VRRE:** Reasoning process quality recognition
- **Impact:** Encourages logical thinking over answer memorization

Integrated Multi-Domain Testing

- **Current:** Isolated domain benchmarks (GSM8K, ARC, etc.)
- **VRRE:** Cross-domain reasoning consistency assessment
- **Impact:** Reveals hidden domain imbalances

For Practical AI Deployment

Efficiency-Guided Model Selection

Use Case	VRRE Recommendation	Traditional Choice
Edge reasoning	Qwen2:0.5b (152.3 quality/GB)	Larger model for "better" scores
Math applications	Phi-3.5 (73.3% math)	General high-scoring model
Balanced reasoning	Mistral:7b	Same, but efficiency missed
Conversational AI	Apollo variants	Same, reasoning gaps missed

Real-World Performance Prediction

VRRE's speed-quality metrics provide actionable deployment guidance:

- **Quality/Second ratios** for inference planning
- **Domain-specific strengths** for application matching
- **Extraction reliability** for production confidence

For Academic Research

Novel Research Directions Enabled

1. **Architectural Efficiency Studies:** Why does Qwen2:0.5b achieve such efficiency?
2. **Domain Transfer Research:** How to balance math vs logic reasoning?
3. **Training Objective Research:** Optimizing for reasoning vs conversation
4. **Semantic Evaluation Theory:** Beyond format compliance paradigms

Reproducible Reasoning Assessment

VRRE provides standardized semantic evaluation enabling:

- **Cross-study comparisons** with consistent reasoning metrics
- **Incremental improvement tracking** with partial credit systems
- **Architecture analysis** across diverse model types

Technical Innovation

Intelligent Response Parsing

VRRE's core innovation lies in understanding *what* a model is trying to communicate rather than *how* it formats the response. This semantic approach reveals reasoning capabilities that format-dependent benchmarks obscure.

Partial Credit Algorithms

The framework recognizes that reasoning is a process, not just an outcome. Models demonstrating logical thinking—even with incorrect conclusions—receive appropriate credit, encouraging reasoning development over answer memorization.

Confidence Calibration

Each extracted answer includes a confidence score based on:

- Pattern strength in the response

- Presence of reasoning indicators
- Consistency across response segments

This enables reliability assessment and error analysis.

Limitations and Future Work

Current Limitations

1. **Model Dependency:** Requires conversational models capable of explanation
2. **Domain Specificity:** Optimized for reasoning tasks, not general capabilities
3. **Pattern Reliance:** May miss novel reasoning expressions

Future Directions

1. **Multi-Modal Extension:** Vision and audio reasoning assessment
2. **Cross-Language Support:** Semantic parsing in multiple languages
3. **Real-Time Applications:** Live reasoning quality monitoring
4. **Automated Task Generation:** Dynamic reasoning challenge creation

Conclusion

The VANTA Research Reasoning Evaluation (VRRE) represents a fundamental paradigm shift from format-dependent to semantic-based LLM assessment. Our comprehensive multi-model validation across five diverse architectures reveals critical gaps in current evaluation methodology and provides actionable insights for AI development and deployment.

Key Contributions

1. **Benchmark Limitation Exposure:** Demonstrated that standard benchmarks miss 2.5x reasoning improvements and fail to detect severe domain imbalances (50% math vs logic gaps in Phi-3.5)
2. **Efficiency Revolution:** Revealed that Qwen2:0.5b achieves competitive reasoning at 12.5x better efficiency than Mistral:7b, challenging parameter-count assumptions
3. **Training Mismatch Detection:** Identified that conversational optimization doesn't translate to logical reasoning capability, exposing fundamental training objective misalignment
4. **Practical Deployment Guidance:** Provided speed-quality trade-off analysis enabling evidence-based model selection for real-world applications

Research Impact

VRRE's semantic understanding approach solves fundamental evaluation problems:

- **Format Independence:** 100% extraction success across diverse architectures
- **Process Recognition:** Partial credit for reasoning quality, not just final answers
- **Cross-Domain Assessment:** Reveals reasoning consistency patterns invisible to specialized benchmarks
- **Efficiency Integration:** Practical deployment metrics beyond pure accuracy

Broader Implications

This framework enables a new generation of AI research focused on:

- **Architectural efficiency** over parameter scaling
- **Domain-balanced training** preventing reasoning gaps
- **Semantic evaluation** beyond format compliance
- **Real-world performance** prediction and optimization

VRRE's open-source availability facilitates immediate adoption across the research community, advancing the field toward more meaningful and practical AI evaluation. The framework's ability to detect improvements invisible to standard benchmarks makes it essential for responsible AI development and deployment.

Future Research Enabled

Our findings open multiple research directions:

- Investigation of Qwen2's architectural efficiency principles
- Development of domain-balanced reasoning training methods
- Extension of semantic evaluation to multimodal reasoning tasks
- Integration of reasoning quality metrics into training pipelines

VRRE represents not just a new evaluation tool, but a foundation for understanding and improving AI reasoning capabilities in ways that traditional benchmarks cannot achieve.

Code Availability

VRRE is available as open-source software:

- **Repository:** <https://github.com/vanta-research/vrre>
- **License:** Apache 2.0
- **Documentation:** Full API documentation and examples included

Citation

```
@software{vrre2025,  
  title={VANTA Research Reasoning Evaluation (VRRE): A Semantic  
Understanding Framework for LLM Reasoning Assessment},  
  author={VANTA Research},  
  year={2025},  
  url={https://github.com/vanta-research/vrre},  
  version={1.0}  
}
```

VANTA Research - Advancing AI through rigorous evaluation methodologies