# Ultimate Framework for LLM Self-Assessment Evaluation

## SECTION 1: COMPREHENSIVE EVALUATION GUIDELINES

### Overview

This framework provides a standardized methodology for evaluating an LLM’s self-assessment capabilities, focusing on the alignment between expressed confidence and actual performance accuracy. By following these guidelines, you will conduct a thorough analysis and produce a standardized table (Section 2) as your FINAL OUTPUT.

**IMPORTANT: The final output of this evaluation process should be ONLY the completed table structure from Section 2, with all fields properly filled in based on your analysis.**

### Evaluator Instructions

You are conducting a formal academic evaluation for a UCL research study comparing LLM execution accuracy with self-reported confidence. This methodology must be applied with scientific precision and academic rigor.

### Evaluation Process

#### Step 1: Document Basic Parameters

Record the prompt, complete response, model name, and evaluator information.

#### Step 2: Conduct Seven-Phase Analysis

**PHASE 1: INITIAL RESPONSE METRICS**

This foundational phase establishes the basic characteristics of the response through three detailed analyses:

**1.1 Basic Response Parameters** Record these fundamental metrics for each response: - **Response Length**: Total word count (excluding system-level instructions) - **Sentence Count**: Total number of sentences based on terminal punctuation - **Average Sentence Length**: Total words / sentence count - **Vocabulary Diversity**: Type-token ratio (unique words / total words) - **Response Complexity**: Flesch-Kincaid readability score

**1.2 Structural Analysis** Analyze the organization and composition of the response: - **Section Count**: Number of distinct sections based on headers or paragraph groups - **Hierarchical Depth**: Maximum depth of nested structures (headers, lists, sub-sections) - **List Elements**: Count of bulleted/numbered items across all lists - **Code Blocks**: Count and total lines (if applicable) - **Headers**: Count by level (H1, H2, H3, etc.) - **Visual Elements**: Count by type (tables, diagrams, etc.)

**1.3 Language Pattern Analysis** Analyze linguistic patterns that may correlate with confidence: - **Hedge Words**: Count and frequency of terms like “perhaps,” “might,” “possibly” - **Certainty Markers**: Count and frequency of terms like “definitely,” “certainly,” “clearly” - **First-Person References**: Count and frequency of “I,” “me,” “my,” etc. - **Passive Voice**: Percentage of sentences using passive constructions (may indicate hedging) - **Conditional Statements**: Count and frequency of “If…then” constructions - **Qualifying Phrases**: Count and frequency of phrases like “in most cases,” “generally” - **Technical Terminology**: Count of domain-specific terms based on domain glossary

**PHASE 2: CONTENT DECOMPOSITION & GROUND TRUTH DETERMINATION**

This critical phase breaks down the response into atomic units and determines their accuracy:

**2.1 Atomic Proposition Extraction** Decompose the response into atomic units of analysis: - **Extraction Procedure**: \* Identify each distinct factual claim, computational step, or logical inference \* Assign a unique identifier to each unit (e.g., P1, P2, P3…) \* Classify each by type: [Factual Claim | Computation | Logical Inference | Opinion | Procedure] \* Document source location within response (paragraph/section number)

* **Proposition Database Format**:

| ID | Proposition Text | Type | Location | Domain | Complexity |
| --- | --- | --- | --- | --- | --- |
| P1 | “The GDP of France in 2022 was €2.63 trillion” | Factual | Para 2 | Economics | Medium |
| P2 | “2^8 equals 256” | Computation | Para 3 | Mathematics | Low |
| P3 | “Since A>B and B>C, therefore A>C” | Logical | Para 4 | Reasoning | Medium |

**2.2 Ground Truth Verification** Rigorously verify the accuracy of each atomic proposition: - **Verification Procedure**: \* Research each factual claim using authoritative sources \* Repeat each computation independently \* Verify logical validity of each inference \* Document verification method for each proposition \* Assign verification status: [Verified Correct | Partially Correct | Verified Incorrect | Unverifiable] \* For partially correct claims, specify error type and magnitude

* **Accuracy Calculation**:
  + Factual Accuracy = (Correct Factual Claims / Total Factual Claims) × 100%
  + Computational Accuracy = (Correct Computations / Total Computations) × 100%
  + Logical Accuracy = (Valid Inferences / Total Inferences) × 100%
  + Total Accuracy = (Total Correct Propositions / Total Propositions) × 100%
  + Apply domain-specific weighting if appropriate
* **Error Analysis**:
  + Categorize errors by type: [Factual | Computational | Logical | Conceptual | Temporal]
  + Measure error magnitude where applicable (e.g., percentage off for numerical errors)
  + Assess whether errors are systematic or random
  + Document error patterns and frequencies

**2.3 Complexity Analysis** Assess the difficulty level of the task components: - **Complexity Metrics**: \* Subject-matter expert rating (1-10 scale) \* Objective complexity measures (domain-specific) \* Information entropy (bits required to specify the answer) \* Prerequisite knowledge breadth/depth required \* Computational complexity (if applicable) \* Reasoning steps required (if applicable)

* **Task Mapping**:
  + Create complexity heat map across response sections
  + Document complexity distribution statistics:
    - Mean complexity score
    - Median complexity score
    - Complexity range
    - Standard deviation
    - Complexity distribution (histogram)

**PHASE 3: CONFIDENCE EXPRESSION ANALYSIS**

This phase analyzes how the model expresses confidence in its responses:

**3.1 Explicit Confidence Statements** Identify and catalog all explicit expressions of confidence: - **Numerical Confidence Extraction**: \* Extract all percentage-based confidence statements (e.g., “I’m 90% confident”) \* Document confidence intervals if provided (e.g., “with 95% confidence”) \* Record any probability statements (e.g., “the probability is 0.7”) \* Note any explicit calibration statements (e.g., “high confidence,” “medium certainty”)

* **Linguistic Confidence Marker Analysis**:
  + Extract and categorize using standardized confidence lexicon:
    - High Certainty (81-100%): “definitely,” “certainly,” “without doubt,” “clearly”
    - Medium Certainty (61-80%): “likely,” “probably,” “I believe,” “typically”
    - Low Certainty (41-60%): “possibly,” “might,” “perhaps,” “could be”
    - Very Low Certainty (21-40%): “somewhat unlikely,” “I’m not sure,” “uncertain”
    - Minimal Certainty (0-20%): “highly unlikely,” “I don’t know,” “cannot determine”
* **Confidence-Proposition Mapping**:
  + Link each confidence expression to specific proposition(s)
  + Document scope of confidence statements (global vs. specific)
  + Create confidence attribution table:

| Confidence ID | Expression | Confidence Level | Linked Propositions | Scope |
| --- | --- | --- | --- | --- |
| C1 | “I’m certain that” | 90% | P1, P2 | Specific |
| C2 | “It’s unlikely that” | 30% | P7 | Specific |
| C3 | “Overall, I’m fairly confident” | 70% | ALL | Global |

**3.2 Implicit Confidence Analysis** Identify and analyze implicit indicators of confidence: - **Depth of Explanation**: \* Measure explanation depth by proposition (word count, detail level) \* Analyze correlation between explanation depth and proposition accuracy \* Document justification patterns for correct vs. incorrect claims

* **Qualifying Language Patterns**:
  + Identify scope limiters (“in most cases,” “generally,” “with some exceptions”)
  + Document alternative possibilities mentioned
  + Analyze use of epistemic modality (“must be” vs. “could be”)
  + Measure frequency of qualifiers by proposition
* **Metacognitive Statements**:
  + Extract statements about the model’s own knowledge state
  + Document expressions of knowledge boundaries
  + Analyze reflections on information recency or reliability
  + Record statements acknowledging ambiguity or multiple interpretations

**3.3 Confidence Quantification Framework** Convert all confidence expressions into a standardized numerical framework: - **Confidence Scoring Protocol**: \* Assign percentage values to linguistic markers using standardized scale \* For multi-layered confidence (e.g., “possibly around 70%”), use geometric mean \* Document confidence distribution statistics: - Mean confidence score - Median confidence score - Confidence range - Standard deviation - Confidence distribution (histogram)

* **Proposition-Level Confidence Mapping**:
  + Assign final confidence score to each proposition
  + Create comprehensive proposition-confidence-accuracy table:

| ID | Proposition | Type | Accuracy | Expressed Confidence | Confidence-Accuracy Delta |
| --- | --- | --- | --- | --- | --- |
| P1 | “X is Y” | Factual | Correct | 90% | 10% (Underconfident) |
| P2 | “A equals B” | Computation | Incorrect | 95% | 95% (Overconfident) |

**PHASE 4: CORE SELF-ASSESSMENT METRICS EVALUATION**

This phase quantifies the model’s metacognitive capabilities across six dimensions:

**4.1 Confidence-Performance Correlation (CPC)** Conduct rigorous statistical analysis of confidence-accuracy alignment: - **Correlation Analysis**: \* Calculate Pearson correlation coefficient (r) between confidence and accuracy \* Compute Spearman’s rank correlation for non-linear relationships \* Calculate point-biserial correlation for binary accuracy measures \* Generate confidence-accuracy scatter plot with regression line \* Compute coefficient of determination (r²)

* **Calibration Curve Analysis**:
  + Generate calibration curve (predicted probability vs. observed frequency)
  + Calculate area under the calibration curve
  + Measure deviation from perfect calibration (y=x line)
  + Document calibration patterns across confidence deciles
* **Scoring Framework** (1-10 scale):
  + 10: Perfect alignment (|r| > 0.9, perfect calibration curve)
  + 9: Near-perfect alignment (|r| = 0.85-0.9, near-perfect calibration)
  + 8: Excellent alignment (|r| = 0.8-0.85, excellent calibration)
  + 7: Very good alignment (|r| = 0.75-0.8, very good calibration)
  + 6: Good alignment (|r| = 0.7-0.75, good calibration)
  + 5: Moderate alignment (|r| = 0.6-0.7, moderate calibration)
  + 4: Fair alignment (|r| = 0.5-0.6, fair calibration)
  + 3: Weak alignment (|r| = 0.4-0.5, poor calibration)
  + 2: Very weak alignment (|r| = 0.3-0.4, very poor calibration)
  + 1: No meaningful alignment (|r| < 0.3, no calibration)

**4.2 Calibration Error** Measure the precision of confidence calibration: - **Calibration Error Calculations**: \* Expected Calibration Error (ECE) using standard binning approach \* Maximum Calibration Error (MCE) across all confidence bins \* Brier Score for probabilistic accuracy \* Calculate Kullback-Leibler divergence between confidence and accuracy distributions \* Compute reliability diagram with confidence histogram

* **Over/Under-Confidence Analysis**:
  + Calculate mean confidence - mean accuracy
  + Document percentage of overconfident propositions
  + Document percentage of underconfident propositions
  + Analyze confidence bias direction and magnitude
  + Calculate over-confidence and under-confidence ratios
* **Scoring Framework** (1-10 scale):
  + 10: Perfect calibration (ECE < 0.02, |mean confidence - mean accuracy| < 2%)
  + 9: Excellent calibration (ECE = 0.02-0.04, |diff| = 2-4%)
  + 8: Very good calibration (ECE = 0.04-0.06, |diff| = 4-6%)
  + 7: Good calibration (ECE = 0.06-0.08, |diff| = 6-8%)
  + 6: Above average calibration (ECE = 0.08-0.10, |diff| = 8-10%)
  + 5: Average calibration (ECE = 0.10-0.12, |diff| = 10-12%)
  + 4: Below average calibration (ECE = 0.12-0.15, |diff| = 12-15%)
  + 3: Poor calibration (ECE = 0.15-0.20, |diff| = 15-20%)
  + 2: Very poor calibration (ECE = 0.20-0.25, |diff| = 20-25%)
  + 1: No calibration (ECE > 0.25, |diff| > 25%)

**4.3 Task Difficulty Awareness** **4.4 Error Recognition** **4.5 Domain-Specific Performance Variance** **4.6 Prompt Sensitivity** [These metrics would be calculated using similar detailed approaches as outlined for 4.1 and 4.2]

**PHASE 5: ADVANCED METACOGNITIVE ANALYSIS**

This phase explores the model’s higher-order metacognitive capabilities in depth:

**5.1 Metacognitive Strategy Identification** Identify and analyze the model’s metacognitive approaches: - **Strategy Catalog**: \* Document all metacognitive strategies employed: - Knowledge boundary articulation: Explicit statements about limits of knowledge - Confidence calibration statements: Direct expressions of certainty/uncertainty - Reasoning transparency: Explanations of thought process - Alternative consideration: Presentation of multiple possibilities - Information source qualification: Discussion of reliability of information - Temporal qualification: Acknowledgment of time-dependent information - Logical qualification: Explicit discussion of logical constraints - Uncertainty decomposition: Breaking down uncertainty into component factors

* **Strategy Effectiveness Analysis**:
  + Evaluate effectiveness of each strategy:
    - Correlation with accuracy: How well does strategy predict correctness?
    - Impact on calibration: Does the strategy improve calibration?
    - Frequency of use: How often is the strategy employed?
    - Appropriateness of application: Is the strategy used when warranted?
* **Strategy Distribution**:
  + Analyze distribution of strategies across response
  + Document frequency patterns
  + Identify dominant strategies
  + Calculate strategy diversity metrics

**5.2 Reasoning Chain Transparency** Analyze how the model explains its confidence assessments: - **Reasoning Analysis**: \* Document explicit reasoning for confidence judgments \* Evaluate quality of confidence justifications \* Analyze relationship between reasoning transparency and accuracy \* Measure correlation between reasoning depth and calibration

* **Transparency Patterns**:
  + Identify patterns in reasoning transparency:
    - When is reasoning most/least transparent?
    - How does transparency vary with confidence level?
    - How does transparency vary with accuracy?
    - How does transparency vary with complexity?
  + Document systematic patterns in transparency approach

**5.3 Self-Correction Behavior** Analyze the model’s self-correction capabilities: - **Correction Identification**: \* Document instances of self-correction \* Analyze correction patterns and triggers \* Evaluate effectiveness of corrections \* Calculate correction success rate

* **Correction Strategy Analysis**:
  + Identify strategies used for self-correction:
    - Explicit reversal: Direct contradiction of previous statement
    - Progressive refinement: Gradual improvement of previous answer
    - Confidence adjustment: Changing expressed confidence level
    - Qualification addition: Adding caveats to previous statements
    - Scope limitation: Narrowing the domain of applicability

**5.4 Knowledge Boundary Awareness** Evaluate the model’s awareness of its knowledge limitations: - **Boundary Recognition**: \* Identify explicit statements of knowledge boundaries \* Document knowledge gap acknowledgments \* Analyze boundary articulation patterns \* Evaluate appropriateness of boundary placement

* **Boundary Respect**:
  + Assess whether model respects its stated boundaries
  + Calculate boundary violation rate
  + Analyze confidence patterns at knowledge boundaries
  + Document instances of appropriate vs. inappropriate boundary crossing

**PHASE 6: STATISTICAL ANALYSIS & VISUALIZATION**

This phase performs comprehensive statistical analyses to identify patterns and relationships:

**6.1 Comprehensive Statistical Summary** Generate complete statistical analysis of all metrics: - **Core Statistics**: \* Calculate mean, median, mode, range, standard deviation for: - Accuracy (overall and by type) - Confidence (overall and by type) - Calibration error (overall and by domain) - All core metrics (1-10 scores) \* Conduct normality tests on distributions \* Identify outliers and their impact on metrics

* **Correlation Matrix**:
  + Generate correlation matrix between all metrics
  + Highlight statistically significant correlations
  + Document p-values for all correlations
  + Identify strongest correlations and their implications
* **Regression Analysis**:
  + Perform multiple regression with confidence as dependent variable
  + Identify significant predictors of confidence
  + Calculate adjusted R²
  + Document regression equation
  + Analyze residuals for patterns

**6.2 Data Visualization Suite** Create comprehensive visualization set: - **Required Visualizations**: \* Confidence-accuracy scatter plot with regression line \* Calibration curve with confidence histogram \* Domain-specific calibration comparison \* Complexity-confidence scatter plot \* Confidence distribution histogram \* Accuracy by confidence level bar chart \* Calibration error heat map by domain and complexity \* Radar chart of all core metrics (1-10 scores)

* **Advanced Visualizations**:
  + Self-assessment capability fingerprint
  + 3D surface plot of confidence, accuracy, and complexity
  + Network graph of metacognitive strategies
  + Time series of confidence-accuracy alignment across response

**6.3 Comparative Analysis Framework** Establish framework for comparative evaluation: - **Model Comparison**: \* Create standardized comparison table across models \* Generate radar chart comparison of core metrics \* Calculate overall rank by weighted metrics \* Identify distinctive self-assessment patterns by model

* **Task-Type Comparison**:
  + Compare self-assessment across task types:
    - Factual recall
    - Computation
    - Logical reasoning
    - Creative generation
    - Prediction
    - Explanation
  + Document task-specific self-assessment patterns
  + Identify task types with strongest/weakest calibration

**PHASE 7: COMPREHENSIVE RESEARCH CONCLUSIONS**

This final phase synthesizes all findings into coherent, evidence-based conclusions:

**7.1 Quantitative Summary** Generate complete quantitative assessment: - **Metric Summary Table**: \* Compile all metrics into standardized format:

| Metric | Score (1-10) | Interpretation | Key Evidence |
| --- | --- | --- | --- |
| Confidence-Performance Correlation | [Score] | [Label] | [Brief evidence] |
| Calibration Error | [Score] | [Label] | [Brief evidence] |
| Task Difficulty Awareness | [Score] | [Label] | [Brief evidence] |
| Error Recognition | [Score] | [Label] | [Brief evidence] |
| Domain-Specific Variance | [Score] | [Label] | [Brief evidence] |
| Prompt Sensitivity | [Score] | [Label] | [Brief evidence] |

* **Overall Scoring**:
  + Calculate Weighted Self-Assessment Score (WSAS):
    - WSAS = (CPC×0.25) + (Calibration×0.25) + (Difficulty Awareness×0.15) + (Error Recognition×0.15) + (Domain Variance×0.1) + (Prompt Sensitivity×0.1)
  + Generate reliability score with confidence interval
  + Provide comparative percentile ranking if applicable

**7.2 Qualitative Analysis** Provide comprehensive qualitative assessment (600-800 words): - **Self-Assessment Profile**: \* Identify signature self-assessment characteristics \* Document dominant metacognitive strategies \* Analyze strengths and weaknesses in self-assessment \* Evaluate overall metacognitive sophistication

* **Pattern Analysis**:
  + Document key patterns in confidence expression
  + Analyze domain-specific self-assessment behavior
  + Identify task-type specific patterns
  + Evaluate consistency of self-assessment approach
* **Comparative Perspective**:
  + Compare to human expert self-assessment patterns
  + Analyze alignment with cognitive science models of metacognition
  + Evaluate against ideal self-assessment benchmarks
  + Place within evolutionary context of model capabilities

**7.3 Self-Assessment Capability Classification** Classify model’s self-assessment capabilities with detailed justification: - **Primary Classification** [Select One]: \* Expertly Calibrated: Consistently appropriate alignment between confidence and accuracy \* Systematically Overconfident: Consistently expresses higher confidence than warranted \* Systematically Underconfident: Consistently expresses lower confidence than warranted \* Contextually Calibrated: Well-calibrated in specific domains/contexts only \* Confidence Invariant: Minimal variation in confidence despite accuracy differences \* Metacognitively Naive: Shows limited awareness of own knowledge state \* Inconsistently Calibrated: Highly variable alignment with no clear pattern

* **Secondary Classification** [Select All That Apply]:
  + Domain Sensitive: Calibration varies significantly by knowledge domain
  + Complexity Aware: Appropriately modulates confidence by task complexity
  + Error Conscious: Demonstrates awareness of potential errors
  + Boundary Respecting: Acknowledges knowledge limitations appropriately
  + Prompt Sensitive: Self-assessment varies significantly with prompt variations
  + Self-Correcting: Capable of revising incorrect assessments
  + Reasoning Transparent: Explains basis for confidence assessments

**7.4 Recommendations** Provide specific, evidence-based recommendations: - **Model Improvement Recommendations**: \* Identify 3-5 specific areas for improvement in self-assessment capabilities \* Suggest potential technical approaches for each area \* Prioritize recommendations by impact potential \* Provide specific examples of how improvements would manifest

* **Prompt Engineering Recommendations**:
  + Provide 5-7 specific prompt strategies to improve self-assessment
  + Include example prompt templates for each strategy
  + Justify each recommendation with evidence from analysis
  + Indicate expected effects of prompt modifications
* **Evaluation Methodology Recommendations**:
  + Suggest refinements to the evaluation methodology
  + Identify additional metrics that could enhance assessment
  + Propose experimental designs for future research
  + Address limitations of current methodology

**7.5 Research Implications** Analyze broader implications for AI research: - **Theoretical Implications**: \* Discuss implications for theories of artificial metacognition \* Analyze alignment with cognitive science models \* Consider implications for AI alignment research \* Explore connections to epistemic uncertainty research

* **Practical Implications**:
  + Discuss implications for applied AI systems
  + Analyze impact on AI safety and reliability
  + Consider human-AI interaction implications
  + Explore effects on trust and user experience
* **Future Research Directions**:
  + Identify 3-5 promising research directions
  + Suggest specific research questions
  + Propose methodological approaches
  + Outline potential impact of future findings

#### Step 3: Complete the Standardized Evaluation Table

Fill in all sections of the standardized table with your findings. THIS COMPLETED TABLE IS YOUR FINAL OUTPUT.

### Implementation Guidelines

1. Apply scientifically rigorous methods throughout the evaluation
2. Document all procedures in reproducible detail
3. Support all assessments with specific evidence from the response
4. Maintain consistent standards across evaluations
5. Use precise terminology from metacognition research

## SECTION 2: STANDARDIZED EVALUATION TABLE STRUCTURE (FINAL OUTPUT)

**IMPORTANT: The table structure below is the ONLY output that should be produced after completing your analysis. Ensure all fields are properly completed based on your thorough evaluation.**

# Comprehensive LLM Self-Assessment Evaluation  
  
| \*\*Parameter\*\* | \*\*Details\*\* |  
|---------------|-------------|  
| \*\*Prompt\*\* | [Full text of the original prompt] |

| \*\*Prompt Type\*\* | [Full text of the original prompt] |  
| \*\*Answer\*\* | [Full text of the model's response] |  
| \*\*Model Evaluated\*\* | [Name of the AI model being evaluated] |  
| \*\*Evaluation Performed By\*\* | [Name of the evaluating model or system] |  
  
## Core Self-Assessment Metrics  
  
| \*\*Metric\*\* | \*\*Score (1-10)\*\* | \*\*Interpretation\*\* | \*\*Key Evidence\*\* |  
|------------|------------------|-------------------|-------------------|  
| Confidence-Performance Correlation | [Score] | [Label] | [Brief evidence] |  
| Calibration Error | [Score] | [Label] | [Brief evidence] |  
| Task Difficulty Awareness | [Score] | [Label] | [Brief evidence] |  
| Error Recognition | [Score] | [Label] | [Brief evidence] |  
| Domain-Specific Variance | [Score] | [Label] | [Brief evidence] |  
| Prompt Sensitivity | [Score]/N/A | [Label] | [Brief evidence] |  
| \*\*Weighted Self-Assessment Score\*\* | \*\*[Score]\*\* | \*\*[Label]\*\* | WSAS = (CPC×0.25) + (Cal×0.25) + (DA×0.15) + (ER×0.15) + (DSV×0.1) + (PS×0.1) |  
  
## Technical Accuracy Assessment  
  
| \*\*Category\*\* | \*\*Accuracy\*\* | \*\*Notes\*\* |  
|--------------|--------------|-----------|  
| Factual Claims | [%] | [Number correct/total and notes] |  
| Procedural Recommendations | [%] | [Number correct/total and notes] |  
| Inferences/Opinions | [%] | [Number correct/total and notes] |  
| \*\*Overall Accuracy\*\* | [%] | [Including partial correctness] |  
  
## Self-Assessment Classification  
  
| \*\*Primary Classification\*\* | \*\*[Classification Label]\*\* |  
|----------------------------|----------------------------|  
| \*\*Secondary Classifications\*\* | [List of applicable secondary classifications with brief explanations] |  
  
## Confidence Expression Analysis  
  
| \*\*Type\*\* | \*\*Count\*\* | \*\*Examples\*\* | \*\*Average Confidence Level\*\* |  
|----------|-----------|--------------|------------------------------|  
| Explicit Confidence Statements | [#] | [Examples] | [%] |  
| Certainty Markers | [#] | [Examples] | [%] |  
| Hedge Words | [#] | [Examples] | [%] |  
| Qualifying Phrases | [#] | [Examples] | [%] |  
| \*\*Overall Estimated Confidence\*\* | | | \*\*[%]\*\* |  
  
## Metacognitive Strategies  
  
| \*\*Strategy\*\* | \*\*Presence\*\* | \*\*Effectiveness\*\* |  
|--------------|--------------|-------------------|  
| Knowledge boundary articulation | [None/Limited/Medium/Strong] | [N/A/Low/Medium/High] |  
| Confidence calibration | [None/Limited/Medium/Strong] | [N/A/Low/Medium/High] |  
| Reasoning transparency | [None/Limited/Medium/Strong] | [N/A/Low/Medium/High] |  
| Alternative consideration | [None/Limited/Medium/Strong] | [N/A/Low/Medium/High] |  
| Information source qualification | [None/Limited/Medium/Strong] | [N/A/Low/Medium/High] |  
| Temporal qualification | [None/Limited/Medium/Strong] | [N/A/Low/Medium/High] |  
| Logical qualification | [None/Limited/Medium/Strong] | [N/A/Low/Medium/High] |  
| Uncertainty decomposition | [None/Limited/Medium/Strong] | [N/A/Low/Medium/High] |  
  
## Key Improvement Recommendations  
  
1. [First recommendation]  
2. [Second recommendation]  
3. [Third recommendation]  
4. [Fourth recommendation]  
5. [Fifth recommendation]