From Hype to Metrics: Assessing Agentic LLMs for Complex Analysis

Lessons Learned from Evaluation Design for the Agentic AI for Competition Analysis (AACA) Research Project

Dr. Stacey Bailey, Dr. Christy Doran, and Dr. Keith J. Miller

© 2026 THE MITRE CORPORATION. ALL RIGHTS RESERVED. APPROVED FOR PUBLIC RELEASE. DISTRIBUTION UNLIMITED. CASE 26-0429.

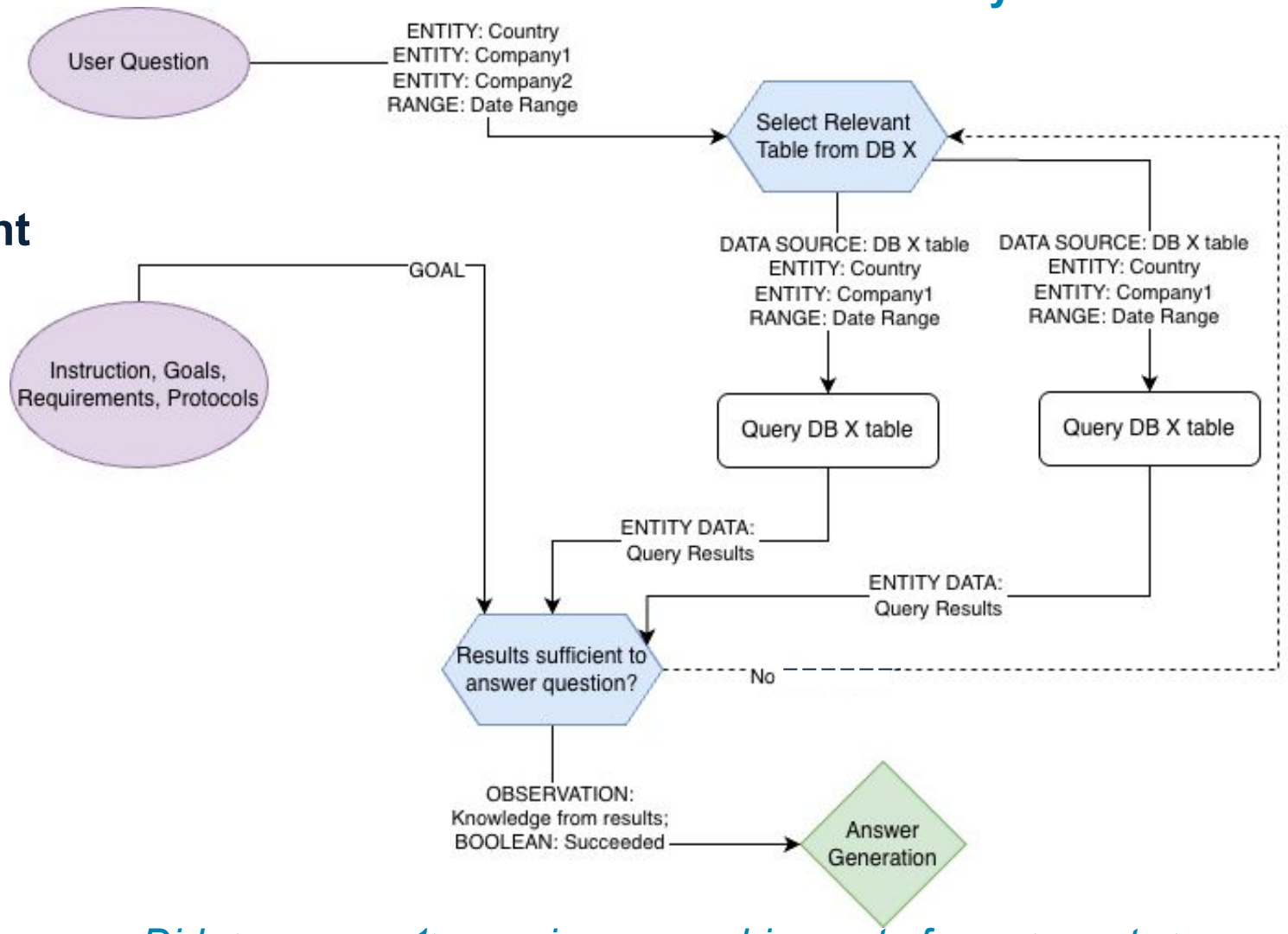
MITRE

Agentic AI for Competition Analysis (AACA)

Question answering to automatically discover, extract, and analyze relevant information from large volumes of structured and unstructured data

- Agentic pipeline
  - LLM-driven tool selection
  - Using LangGraph
- Research goal involved exploring degrees of freedom in terms of
  - Tool utilization
  - Data complexity
  - Complex reasoning power

Notional System Flow



Did <company1> receive more shipments from <country> than <company2> did during <date range>?

MITRE © 2026 THE MITRE CORPORATION. ALL RIGHTS RESERVED.
APPROVED FOR PUBLIC RELEASE. DISTRIBUTION UNLIMITED. CASE 26-0429.

Evaluation Goal and Scope

- **Goal**: To deeply understand the strengths and weaknesses of the end-to-end system to answer the primary research question:

  Can LLMs reliably leverage toolchains for complex analytical reasoning tasks?

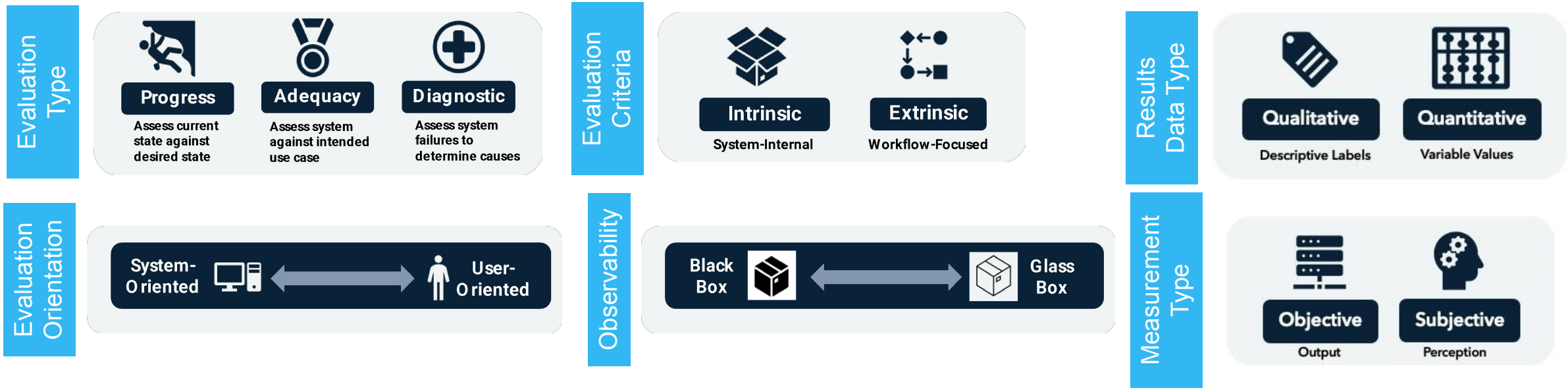
- **Scope**: Deliberately limited to a non-interactive (single-turn) question-answering task
  - Human analytical processes are iterative.
  - Evaluation of conversational interaction between human analyst and the system is a logical next step, but dialogue system evaluation adds significant complexity to the design.

MITRE

© 2026 THE MITRE CORPORATION. ALL RIGHTS RESERVED. APPROVED FOR PUBLIC RELEASE. DISTRIBUTION UNLIMITED. CASE 26-0429.

Evaluation Task Model

Leveraging *Principles of Evaluation in Natural Language Processing* (Paroubek, et al. 2007)



Evaluation Type	Evaluation Orientation	Evaluation Criteria	Observability	Results Data Type	Measurement Type
•Diagnostic •Progress	•System-oriented	•Intrinsic	•Closer to glass box, though some components are black box	•Quantitative	•Objective

© 2026 THE MITRE CORPORATION. ALL RIGHTS RESERVED. APPROVED FOR PUBLIC RELEASE. DISTRIBUTION UNLIMITED. CASE 26-0429.

Additional Evaluation Characteristics

System- and Component-level Evaluation

✅ What is the quality of the final responses produced by the system?
✅ Does the system use the tools correctly?
❌ To what degree do the tools work as intended?

Levels of Guidance (Primary Evaluation Variables)

✅ What kind and how much system guidance is necessary to produce high-quality responses?

MITRE

© 2026 THE MITRE CORPORATION. ALL RIGHTS RESERVED. APPROVED FOR PUBLIC RELEASE. DISTRIBUTION UNLIMITED. CASE 26-0429.

Evaluation Pipeline

© 2026 THE MITRE CORPORATION. ALL RIGHTS RESERVED. APPROVED FOR PUBLIC RELEASE. DISTRIBUTION UNLIMITED. CASE 26-0429.

Eval Corpus and Reference Answers/Tool Calls

13 Information Needs	5-10 Templates Per Need	Filler Entities from Data Sources	Ground Truth for Each Filler Set
▪ Domain questions ▪ Variable task complexity	▪ Specific wording with variables for entities of interest ▪ More complex/simpler wording or syntax ▪ Phrasing that eliminated fillers with default values (e.g., dates)	▪ Fillers present in/across given data sources ▪ Entities known to **not** be present in our data sources (fictitious entities) ▪ Fillers abstracted away from specific named entities (e.g., *in any country* vs. *in Spain*).	▪ Set of correct tool sequences (chains) - required and optional tools and arguments ▪ Set of correct responses for each filler set

Information need: *Understand which of two companies receives more shipments from a given country.*

Question template: Did <company1> receive more shipments from <country> than <company2> did during <date range>?

Filler set: {ABC Company, ACME LLC, France, during the past 5 years}

Instantiated question: Did ABC Company receive more shipments from France than ACME LLC did during the past 5 years?

MITRE

© 2026 THE MITRE CORPORATION. ALL RIGHTS RESERVED. APPROVED FOR PUBLIC RELEASE. DISTRIBUTION UNLIMITED. CASE 26-0429.

Lessons (Re-)Learned in Test Set Development

- Tool (in-)flexibility is a limiting factor for test item variation.
  - Human analysts and LLMs may adapt to subtle differences in wording, but the tools won't.
- Data source representations are a limiting factor for fillers.
  - Selecting fillers that were presented/represented in all relevant data sources was a deliberate simplification, but entity resolution would be necessary in real-world applications
- Time limitations are real.
  - Complete reference answer creation was not possible for all our information needs, but we measured what we could for every information need.
- Automated evaluation of open-ended responses is/has been/always will be a problem for NLP evaluation.
  - No single "correct answer", plus the user prompt can introduce potential ambiguity in what's a correct response.

MITRE

© 2026 THE MITRE CORPORATION. ALL RIGHTS RESERVED. APPROVED FOR PUBLIC RELEASE. DISTRIBUTION UNLIMITED. CASE 26-0429.

Primary Evaluation Variables:

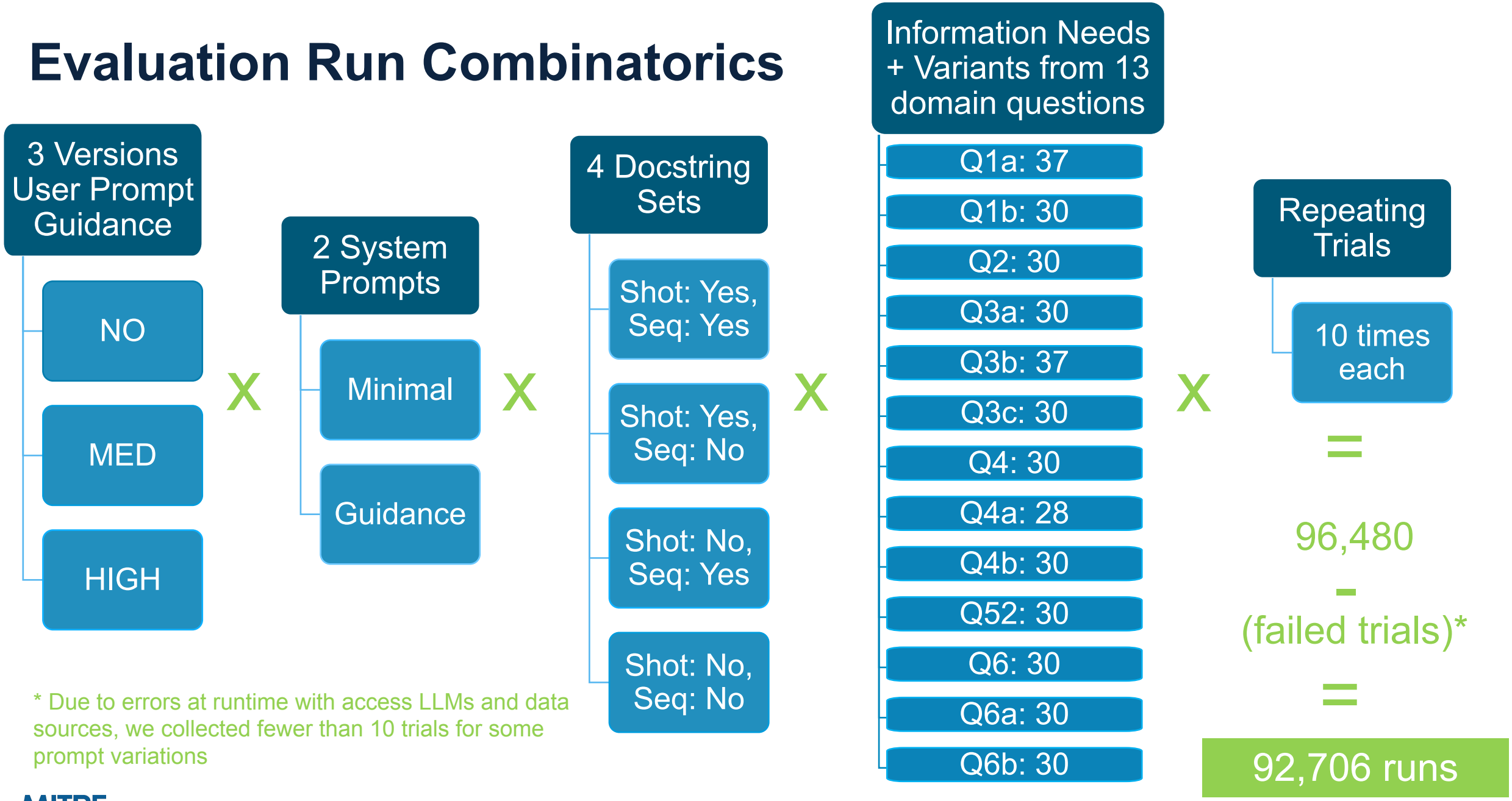
How much instructional micro-management (“hand-holding”) does the system need?

User Prompt Guidance	•No guidance •Medium guidance (data source requirement) •High guidance (tool/task sequencing)	Varies by information need
System Prompt Guidance	•Minimal •Guidance (Correct answer characteristics)	Consistent across information needs
Tool Docstrings	•Few shot examples: Yes, Sequencing: Yes •Few shot examples: Yes, Sequencing: No •Few shot examples: No, Sequencing: Yes •Few shot examples: No, Sequencing: No	Defined for each tool

MITRE

© 2026 THE MITRE CORPORATION. ALL RIGHTS RESERVED. APPROVED FOR PUBLIC RELEASE. DISTRIBUTION UNLIMITED. CASE 26-0429.

Evaluation Run Combinatorics



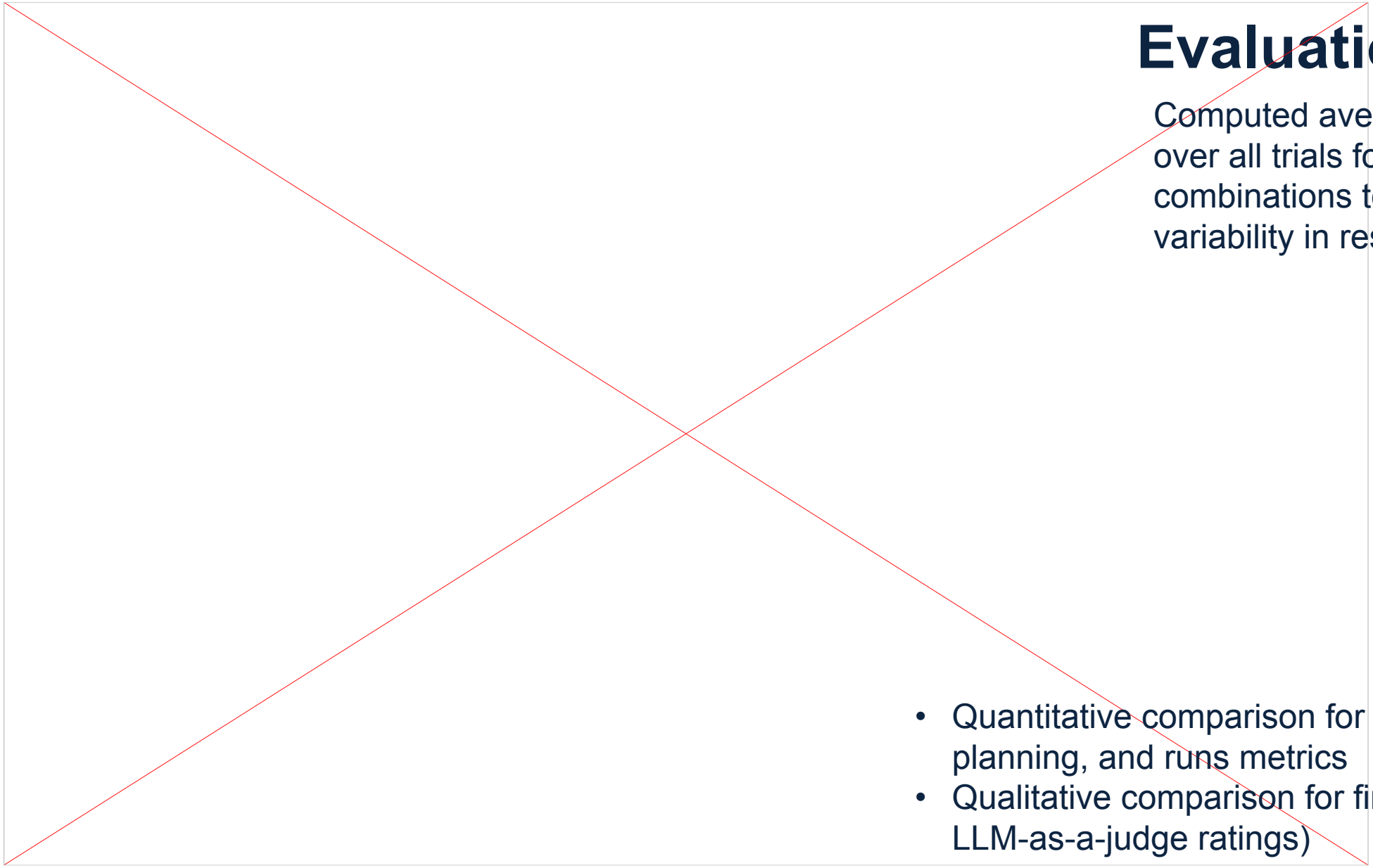
* Due to errors at runtime with access LLMs and data sources, we collected fewer than 10 trials for some prompt variations

MITRE

© 2026 THE MITRE CORPORATION. ALL RIGHTS RESERVED. APPROVED FOR PUBLIC RELEASE. DISTRIBUTION UNLIMITED. CASE 26-0429.

Evaluation Metrics

Computed average performance over all trials for each combinations to account for variability in responses



- Quantitative comparison for tools, steps, planning, and runs metrics
- Qualitative comparison for final responses (via LLM-as-a-judge ratings)

MITRE

© 2026 THE MITRE CORPORATION. ALL RIGHTS RESERVED. APPROVED FOR PUBLIC RELEASE. DISTRIBUTION UNLIMITED. CASE 26-0429.

LLM-as-a-judge for Open-Ended Response Assessment

- Growing practice of using LLM-as-a-judge in evaluation when exhaustive human review of system output is not feasible
  - Adopted this approach for Task Completion and Response Correctness assessment
- Inherent risks require mitigation to provide confidence in quality of assessment:
  - LLM-as-a-judge is still a non-deterministic LLM and may not be consistent in ratings
  - Some human review should be required to calibrate ratings
  - Iteration over LLM-as-a-judge prompts might be required to align ratings to expectations
- Assessing LLM-as-a-judge prompt iterations:
  - Task Completion and Response Correctness have a logical relationship that can be leveraged as a sanity check => Response cannot both be Correct and Incomplete
  - Manual rating of a subset of responses to grade the grader (light meta-evaluation)

© 2026 THE MITRE CORPORATION. ALL RIGHTS RESERVED. APPROVED FOR PUBLIC RELEASE. DISTRIBUTION UNLIMITED. CASE 26-0429.

Other LLM-as-a-judge Observations

- At times, the LLM's rating conflicted with its subsequent explanation of the rating.
  - Need to further assess and characterize the quality and limitations of the LLM-as-a-judge responses.
- Our Correctness prompt compared the system answer to the reference answer and provided characteristics of a good answer as response guidelines.
  - At times, these answer characteristics were more detailed and in conflict with what was in the reference answer.
  - Need for either further revisions of the prompting approach or revision of the reference answer set to capture a fuller range of acceptable answers.
- To automate the metrics collection, we generated the ratings and explanations as structured output, but the LLM did not always produce the expected format.
  - Additional analysis of these conflicts could help to identify the conditions in which this problem arises, and guide improvements in either the evaluation prompts or reference sets.

© 2026 THE MITRE CORPORATION. ALL RIGHTS RESERVED. APPROVED FOR PUBLIC RELEASE. DISTRIBUTION UNLIMITED. CASE 26-0429.

Some results

© 2026 THE MITRE CORPORATION. ALL RIGHTS RESERVED. APPROVED FOR PUBLIC RELEASE. DISTRIBUTION UNLIMITED. CASE 26-0429.

System Hand-holding and Tool Calling

User Guidance	System Prompt Guidance	Docstring Shots	Docstring Sequencing	Average of Tool Call Precision	Average of Tool Call Recall	Average of Tool Call Lev Similarity
HIGH	Yes	No	No	0.93	0.86	0.72
HIGH	No	No	No	0.91	0.84	0.69
HIGH	Yes	Yes	No	0.90	0.78	0.67
HIGH	No	Yes	No	0.88	0.79	0.65
MID	Yes	No	No	0.91	0.80	0.70
MID	No	No	No	0.86	0.76	0.65
MID	Yes	Yes	No	0.88	0.72	0.65
MID	No	Yes	No	0.85	0.73	0.63
NO	Yes	No	No	0.90	0.76	0.69
NO	No	No	No	0.86	0.72	0.63
NO	Yes	Yes	No	0.87	0.70	0.63
NO	No	Yes	No	0.86	0.70	0.62
HIGH	Yes	No	Yes	0.81	0.81	0.62
HIGH	No	No	Yes	0.76	0.82	0.59
HIGH	Yes	Yes	Yes	0.83	0.78	0.63
HIGH	No	Yes	Yes	0.78	0.77	0.58
MID	Yes	No	Yes	0.79	0.74	0.60
MID	No	No	Yes	0.74	0.75	0.56
MID	Yes	Yes	Yes	0.81	0.71	0.60
MID	No	Yes	Yes	0.74	0.70	0.54
NO	Yes	No	Yes	0.75	0.71	0.56
NO	No	No	Yes	0.70	0.69	0.52
NO	Yes	Yes	Yes	0.78	0.67	0.56
NO	No	Yes	Yes	0.73	0.68	0.53

Feature with highest performance 3/3 metrics

Feature with highest performance 2/3 metrics

Highest to lowest scores for P/R/Lev

Results show

- HIGH user guidance helps performance
- Docstring sequencing guidance hurt performance
- Having some system prompt guidance generally helped
- Picture is less clear for docstring shots.

MITRE

© 2026 THE MITRE CORPORATION. ALL RIGHTS RESERVED. APPROVED FOR PUBLIC RELEASE. DISTRIBUTION UNLIMITED. CASE 26-0429.

System Hand-holding and Response Correctness

User Guidance	System Prompt Guidance	Docstring Shots	Docstring Sequencing	Percent Completely Correct
HIGH	Yes	No	No	53.4%
HIGH	Yes	Yes	No	48.9%
HIGH	No	No	No	52.2%
HIGH	No	Yes	No	48.3%
MID	Yes	No	No	46.7%
MID	Yes	Yes	No	43.2%
MID	No	No	No	47.1%
MID	No	Yes	No	42.5%
NO	Yes	No	No	45.9%
NO	Yes	Yes	No	44.3%
NO	No	No	No	43.1%
NO	No	Yes	No	39.6%
HIGH	Yes	No	Yes	45.3%
HIGH	Yes	Yes	Yes	44.1%
HIGH	No	No	Yes	46.3%
HIGH	No	Yes	Yes	41.0%
MID	Yes	No	Yes	40.1%
MID	Yes	Yes	Yes	38.7%
MID	No	No	Yes	39.0%
MID	No	Yes	Yes	37.1%
NO	Yes	No	Yes	39.1%
NO	Yes	Yes	Yes	39.8%
NO	No	No	Yes	36.7%
NO	No	Yes	Yes	35.9%

Feature with highest performance

Highest to lowest scores for % Completely Correct

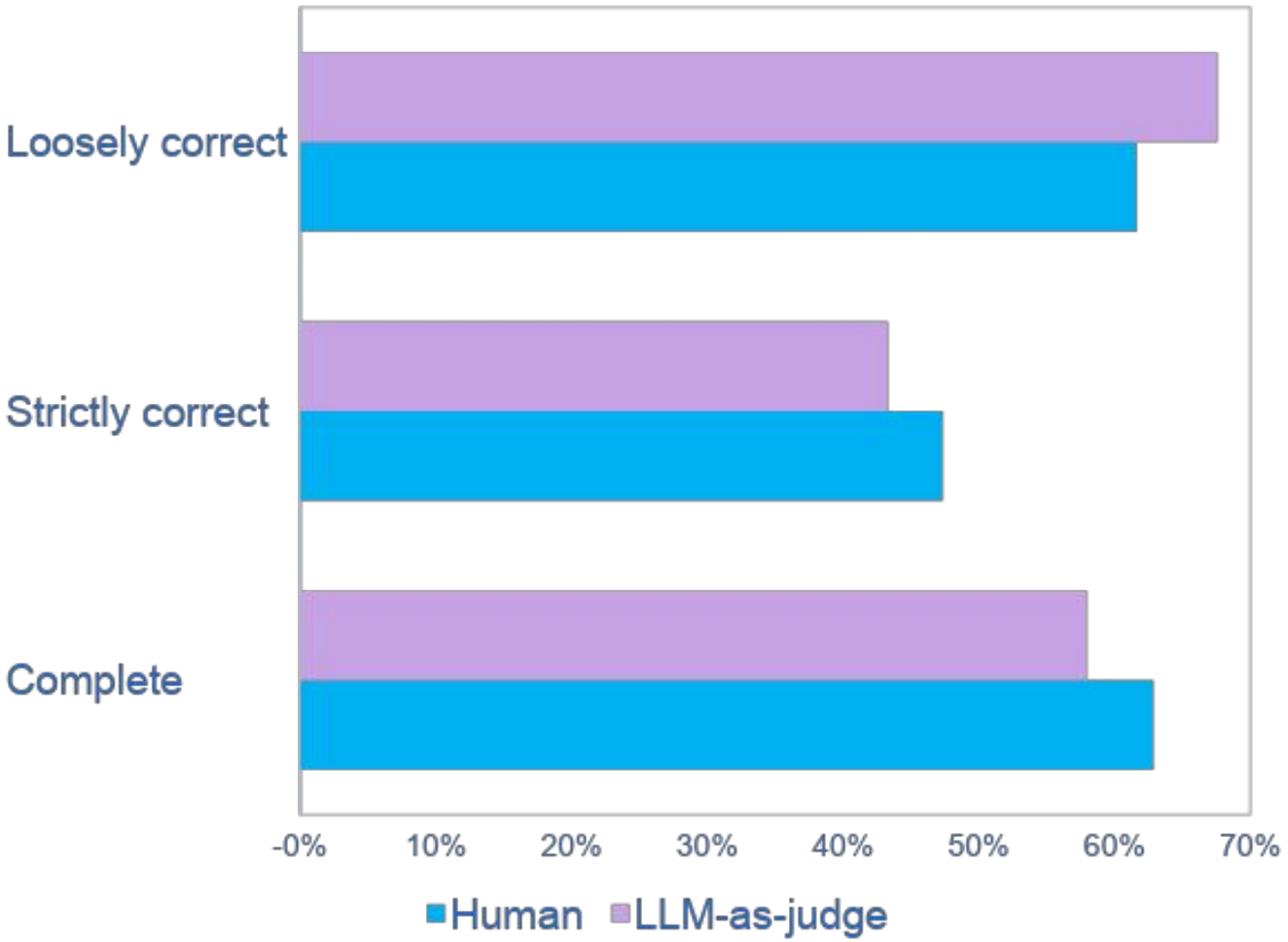
LLM-as-judge response correctness shows a similar pattern:

- HIGH guidance -> higher ratings
- Docstring sequencing -> lower ratings

MITRE

© 2026 THE MITRE CORPORATION. ALL RIGHTS RESERVED. APPROVED FOR PUBLIC RELEASE. DISTRIBUTION UNLIMITED. CASE 26-0429.

LLM vs. Manual Ratings



- Overall, there is about a 5% delta between the Human and LLM-as-judge ratings, with the caveat that there was substantial variability on individual questions.
- Humans judged answers to be complete 4.9% more often.
- The LLM-as-judge was 4% stricter on absolute correctness, but put many more answers in the middle “Partially Correct” category than the human judges did, resulting in a 6% higher loosely correct rating.
- While the aggregate numbers show good general alignment, there is still room for improvement in consistency of alignment at a more granular level.

MITRE

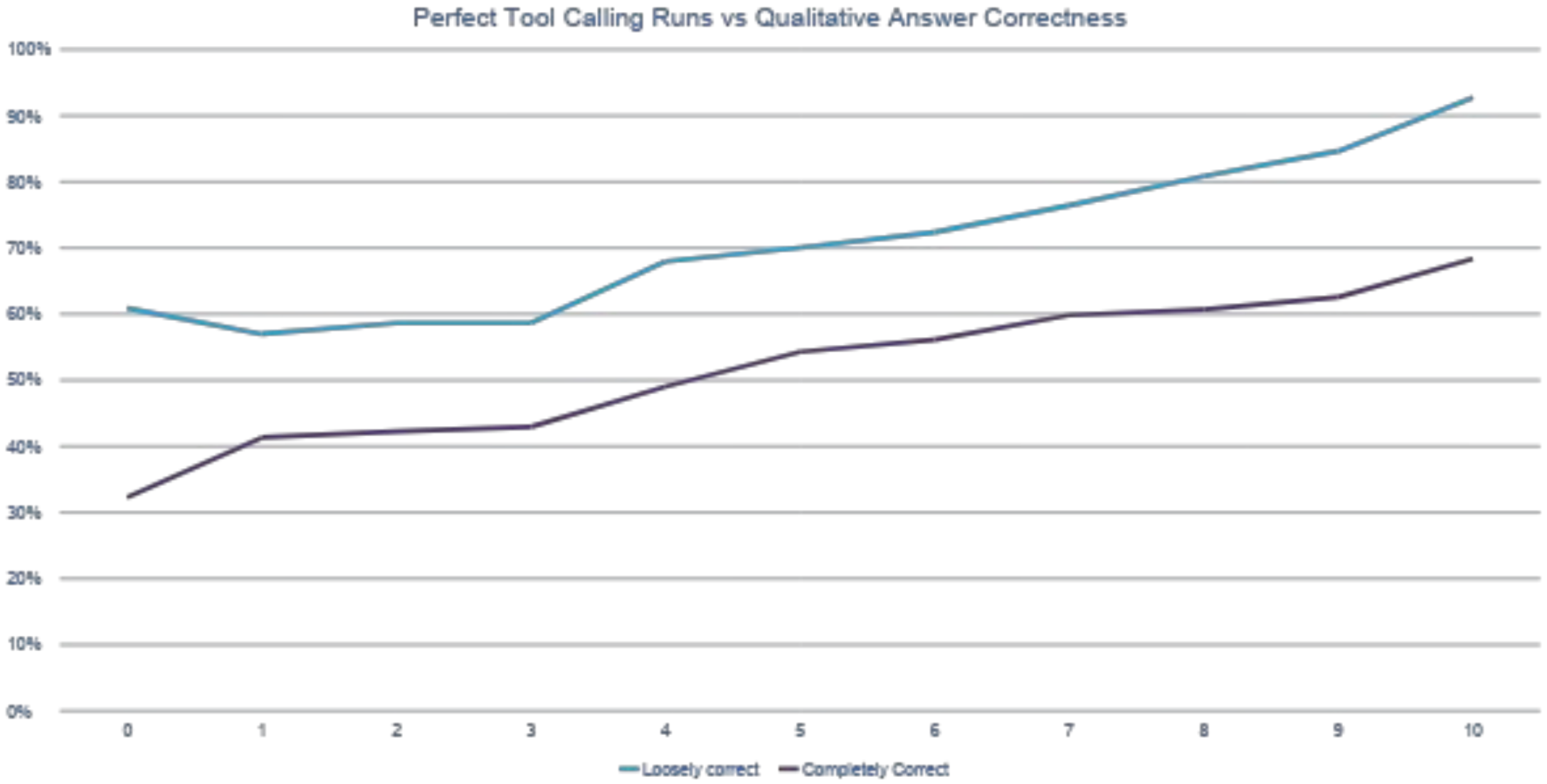
© 2026 THE MITRE CORPORATION. ALL RIGHTS RESERVED. APPROVED FOR PUBLIC RELEASE. DISTRIBUTION UNLIMITED. CASE 26-0429.

Response Consistency

- Non-determinism in LLMs is a feature, not a bug: The same LLM with the same configuration, when given the same prompt, is not guaranteed to produce the same response each time.
- Critical to understand how stable the responses are if using agentic LLMs to support production pipelines
  - Repeating trials (10x) and examining the performance was our approach to investigating this.
- Only scratched the surface of analysis -- Deeper dives needed to understand
  - When the responses were more likely to be consistent
  - Whether question variants (wording and entity types) affected performance or consistency

© 2026 THE MITRE CORPORATION. ALL RIGHTS RESERVED. APPROVED FOR PUBLIC RELEASE. DISTRIBUTION UNLIMITED. CASE 26-0429.

Quantitative + qualitative



MITRE

© 2026 THE MITRE CORPORATION. ALL RIGHTS RESERVED. APPROVED FOR PUBLIC RELEASE. DISTRIBUTION UNLIMITED. CASE 26-0429.

Response Consistency

© 2026 THE MITRE CORPORATION. ALL RIGHTS RESERVED. APPROVED FOR PUBLIC RELEASE. DISTRIBUTION UNLIMITED. CASE 26-0429.

Final Thoughts on Executing LLM Evaluations

- Don’t skimp on tracking and instrumentation.
  - Something always goes wrong, which means trials will need to be re-run, sometimes multiple times. This not only needs to be factored into the evaluation timeline, but it also points to the need for careful, methodical bookkeeping (e.g., experiment naming conventions) and explicit version control to keep track of what has been run or needs to be re-run.
  - We used MLFlow for our experiment tracking.
- Log, log, log.
- What you can measure isn’t always what you wanted.
  - What metrics are appropriate depends on your evaluation goals, the task(s) being evaluated, and the data available. All measures and metrics have strengths and weakness, so there is no single metric that is best for every use case even for the same task.
  - Do the best you can with the time and resources you have.

MITRE

© 2026 THE MITRE CORPORATION. ALL RIGHTS RESERVED. APPROVED FOR PUBLIC RELEASE. DISTRIBUTION UNLIMITED. CASE 26-0429.

Questions?

© 2026 THE MITRE CORPORATION. ALL RIGHTS RESERVED. APPROVED FOR PUBLIC RELEASE. DISTRIBUTION UNLIMITED. CASE 26-0429.

Backups

© 2026 THE MITRE CORPORATION. ALL RIGHTS RESERVED. APPROVED FOR PUBLIC RELEASE. DISTRIBUTION UNLIMITED. CASE 26-0429.

Task Complexity

- How to define complexity in a measurable way that discriminates among human analytical QA tasks
  - Multiple indicators
  - Empirically determine if any correlate to performance
- Not all indicators have different values across our questions
  - Non-discriminating indicators might still be relevant for other types of analytical questions or tasks.

MITRE

© 2026 THE MITRE CORPORATION. ALL RIGHTS RESERVED. APPROVED FOR PUBLIC RELEASE. DISTRIBUTION UNLIMITED. CASE 26-0429.

Performance and Task Complexity

- “Countable” indicators (e.g, number of steps, number of inputs) showed no signal
- Some signal in comparing the supply chain question performance against the technology landscape questions:
  - Overall supply chain performance 62% Correct vs technology landscape overall 52% Correct
  - Supply chain questions patterned together for multiple complexity indicators; technology landscape questions also patterned together.
- More work required to tease apart what might be happening

MITRE

© 2026 THE MITRE CORPORATION. ALL RIGHTS RESERVED. APPROVED FOR PUBLIC RELEASE. DISTRIBUTION UNLIMITED. CASE 26-0429.