Pacific Northwest
NATIONAL LABORATORY

PNNL-40099

What Makes an LLM Agent Team Work? Evaluating Teams Through Randomized-Block Tournaments

Multi-Agent Randomized-Block LLM Ensembles (MARBLE) for Reactor Diagnostics

September 2026

Kefei Wang
Kaustav Bhattacharjee
Gregory Roek
Ted Fujimoto

U.S. DEPARTMENT of ENERGY

Prepared for the U.S. Department of Energy
under Contract DE-AC05-76RL01830

DISCLAIMER

This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor any agency thereof, nor Battelle Memorial Institute, nor any of their employees, makes **any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights**. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof, or Battelle Memorial Institute. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof.

PACIFIC NORTHWEST NATIONAL LABORATORY
operated by
BATTELLE
for the
UNITED STATES DEPARTMENT OF ENERGY
under Contract DE-AC05-76RL01830

Printed in the United States of America

Available to DOE and DOE contractors from
the Office of Scientific and Technical Information,
P.O. Box 62, Oak Ridge, TN 37831-0062
www.osti.gov
ph: (865) 576-8401
fax: (865) 576-5728
email: reports@osti.gov

Available to the public from the National Technical Information Service
5301 Shawnee Rd., Alexandria, VA 22312
ph: (800) 553-NTIS (6847)
or (703) 605-6000
email: info@ntis.gov
Online ordering: http://www.ntis.gov

PNNL-40099

What Makes an LLM Agent Team Work? Evaluating Teams Through Randomized-Block Tournaments

Multi-Agent Randomized-Block LLM Ensembles (MARBLE) for Reactor Diagnostics

September 2026

Kefei Wang
Kaustav Bhattacharjee
Gregory Roek
Ted Fujimoto

Prepared for
the U.S. Department of Energy
under Contract DE-AC05-76RL01830

Pacific Northwest National Laboratory
Richland, Washington 99354

Abstract

Understanding what makes an LLM agent team effective requires systematically evaluating how individual agent configurations influence overall team performance. Multi-Agent Randomized-Block LLM Ensembles (MARBLE) developed a framework for evaluating role-specific model, prompt, and domain-knowledge choices in a technical diagnostic workflow. The project built a reusable agent stack for Generic Pressurized Water Reactor (GPWR) simulator data and a randomized-block tournament method for comparing agent teams under matched conditions. A live FY2026 prompt-team tournament evaluated the two-role Rapid workflow with one fixed candidate model, one Skill preset, and one independent evaluator. It completed 32 candidate executions across three rounds, observed eight of nine possible prompt teams, used about 437,000 tokens, and recorded no execution failures. The leading observed team paired a coverage-focused prompt in the Rapid Diagnostician with an evidence-calibration prompt in the Evidence Adjudicator and reached a mean objective score of 0.924. Twenty-nine of the 32 reports received the evaluator's maximum evidence-quality score, limiting the rubric's ability to distinguish among teams. The study therefore supports a narrow but important conclusion: role-specific prompt assignments influenced observed team performance and token use, although the study did not establish diagnostic accuracy. The project also implemented a blinded, fingerprinted Sequential RBTT lifecycle for future ground-truth evaluation.

Summary

Understanding what makes an LLM agent team effective requires examining how different configurations influence overall team performance. A workflow assigns models, prompts, knowledge, and responsibilities to different roles, yet teams are often assembled from intuition or a single demonstration. That practice gives decision makers little evidence about whether another role assignment would produce a clearer diagnosis, use fewer tokens, or fail less often. MARBLE addresses this challenge through a randomized-block tournament framework for systematically comparing agent teams.

The project delivered three related programs. The reactor demo runs Rapid, Sequential, and Parallel diagnostic workflows. The prompt-team RBTT program searches short, role-specific prompt treatments while holding the workflow, data, Skill preset, model, and evaluator fixed. The static Sequential RBTT program compares model-and-Skill Candidates on blinded Task Blocks under a frozen Experiment Fingerprint and an append-only event log. The programs share data adapters and agent implementations, but each answers a different research question.

The live Rapid tournament showed that the same instruction did not behave the same way in both roles. Conservative triage had the highest sampled mean when assigned to the Rapid Diagnostician, while evidence calibration had the highest sampled mean in the Evidence Adjudicator. The leading exact team was observed twice, and the full objective spread was 0.075. These results identify a promising assignment; the small sample and saturated judge scores do not support a superiority or statistical-significance claim.

The main remaining evidence gap is diagnostic correctness. The live evaluator reviewed calibration, traceability, alternatives, and limitations without seeing hidden fault labels. The follow-on Sequential protocol addresses that gap with a deterministic Diagnosis Score, blinded six-case Task Blocks, balanced schedules, and held-out confirmation. The software implements that lifecycle and passes the repository's 187-test default suite. Paid live Development and Pilot Experiments remain future research activities.

Acknowledgments

This research was supported by the FY2026 Mission Seed under the Laboratory Directed Research and Development Program at Pacific Northwest National Laboratory. Pacific Northwest National Laboratory is operated for the U.S. Department of Energy by Battelle Memorial Institute under Contract No. DE-AC05-76RL01830.

Acronyms and Abbreviations

AFW	auxiliary feedwater
CWS	circulating water system
DOE	U.S. Department of Energy
GPWR	Generic Pressurized Water Reactor
LDRD	Laboratory Directed Research and Development
LLM	large language model
MARBLE	Multi-Agent Randomized-Block LLM Ensembles
MFW	main feedwater system
PNNL	Pacific Northwest National Laboratory
PWR	pressurized-water reactor
RBTT	Randomized-Block Team Tournament
RCS	reactor coolant system
SG	steam generator

Contents

Abstract ii
Summary iii
Acknowledgments iv
Acronyms and Abbreviations v
Introduction 1
Data and Diagnostic Task 2
Algorithm 4
Software Architecture 6
Results 8
Conclusions and Next Steps 10
References 11
Appendix A Reproducibility Record A.1

Figures

Figure 1. Prompt-team RBTT within a fixed diagnostic workflow 4
Figure 2. GPWR multi-agent diagnostic workflows 6

Tables

Table 1. GPWR dataset used by MARBLE 2
Table 2. Required diagnostic fields and boundaries 3
Table 3. Program boundaries in the MARBLE repository 7
Table 4. Live Rapid prompt-team leaderboard 8

1.0 Introduction

1.1 Background

Large language models can divide a technical task among specialized roles, but the quality of the resulting workflow depends on decisions that are easy to hide. A developer chooses which model fills each role, what instructions it receives, what domain knowledge enters its context, and how intermediate outputs move between roles. A successful demonstration does not show whether those choices were efficient or whether another assignment would perform better under the same conditions.

MARBLE addressed that comparison problem with Randomized-Block Team Tournaments. The method forms role-compatible teams, evaluates them on matched tasks, records resource use and failures, and ranks directly observed teams under a declared rule. The GPWR simulator dataset supplied a technical test bed with many coupled signals, subtle secondary-side faults, and a clear need for calibrated claims.

1.2 Why This Work Matters

Plant operators, digital-twin teams, and program managers need more than a plausible model response when they assess an LLM assistant. They need to know what data every candidate saw, which factors changed between candidates, how the evaluator reached its score, and what the run cost. Matched comparisons also matter because live model outputs can vary from call to call. If teams receive different tasks or evaluators, the ranking mixes the team effect with changes in the test conditions.

The nuclear setting raises the standard for claim discipline. The dataset contains no row-level fault flag, and several CWS and MFW faults move related signals in similar directions. MARBLE therefore requires evidence, alternative explanations, confounders, and stated data limits. Its reports provide diagnostic evidence for research; they do not issue licensed operator instructions or make claims about safety margin.

1.3 Project Objectives and Scope

FY2026 work pursued two connected objectives. The first was a reusable multi-agent diagnostic stack that could compare Rapid, Sequential, and Parallel workflows without embedding GPWR facts in the agent code. The second was an experiment method that could search role assignments within a fixed budget and retain enough evidence to reproduce the result.

To support these objectives,the completed software supports three programs: a workflow demo, a prompt-treatment tournament, and a blinded static Sequential tournament. The live result in this report comes from the prompt-treatment program. The static Sequential lifecycle has simulated validation and has not yet produced a paid live diagnostic-accuracy result.

2.0 Data and Diagnostic Task

2.1 GPWR Simulator Dataset

The local GPWR dataset contains 539 independent CSV trajectories from a generic two-loop pressurized-water reactor simulator. Sixty trajectories represent normal operation and 479 contain fault injections. Each file contains 1,037 samples at 1 Hz across 239 signals. The files total about 1.07 GB and include secondary-side circulating-water and main-feedwater leak or pump-degradation scenarios. One expected file, MFW_leak_v1_06.csv, is absent from the otherwise 540-file grid (MARBLE Project 2026a).

Normal operation spans changing load and environmental conditions. Across the 60 normal runs, reactor thermal power ranges from 2,989.29 to 3,983.12 MW, and circulating-water inlet temperature ranges from 8.38 to 29.33 C. A single global mean would therefore confuse normal operating changes with faults. The data adapter builds a bounded prompt and supplies a normal reference so that the agents can compare signals in operating context.

Table 1. GPWR dataset used by MARBLE

Item	Recorded value
Trajectories	539 total: 60 normal and 479 fault-injected
Trajectory length	1,037 samples from 10:00.000 to 27:16.000
Sampling and variables	1 Hz across 239 process signals
Fault scope	CWS and MFW leaks and pump degradation
Plant ratings	3,983 MW thermal and 1,400 MW electric
Labels	Filename-derived category and timing; no row-level fault flag

2.2 Diagnostic Output and Guardrails

Each case produces a structured diagnosis. The report classifies the trajectory as normal, anomalous, or uncertain and supplies a probability. It then identifies the subsystem and fault mode when the evidence supports them. Severity and onset remain optional because the available labels and signals do not always support precise estimates. The final report also records evidence, credible alternatives, confounders, and limitations.

Table 2. Required diagnostic fields and boundaries

Field	Allowed output	Boundary
Anomaly	NORMAL, ANOMALOUS, UNCERTAIN	Probability required
Subsystem	CWS, MFW, NONE, UNKNOWN	Cited signals required
Fault mode	LEAK, PUMP DEGRADATION, NONE, UNKNOWN	State alternatives and conflicts
Severity and onset	Only when supported	Avoid unsupported precision
Safety and action	Diagnostic evidence	No safety claim or operator command

3.0 Algorithm

3.1 Prompt Team Randomized Block Tournament

The prompt-team tournament changes one short treatment for each active role and freezes the other inputs. The treatment bank contains coverage, evidence calibration, and conservative triage. The Rapid workflow has two active roles and therefore nine possible assignments. Sequential and Parallel each have four active roles and 81 possible assignments. A run with all workflows creates three independent tournaments; it does not select one workflow across the three topologies.

Each tournament round samples role-compatible teams with replacement. To ensure matched comparisons, every unique team in a block receives the same prepared input, workflow, Skill preset, candidate-model assignment, and evaluator. Repeated trials contribute separately to the team's mean. The scheduler carries the leading observed teams into later rounds so that promising assignments receive more evidence. Ranking uses the mean objective because carried teams intentionally receive more trials than teams seen once.

The FY2026 prompt-team objective assigned 80 percent to evaluator quality, 15 percent to token efficiency within the matched block, and 5 percent to successful completion. The JSON result retains every prompt assignment, raw run, judge explanation, token count, score component, failure, and leaderboard entry. This weighted objective belongs to the prompt-treatment program only.

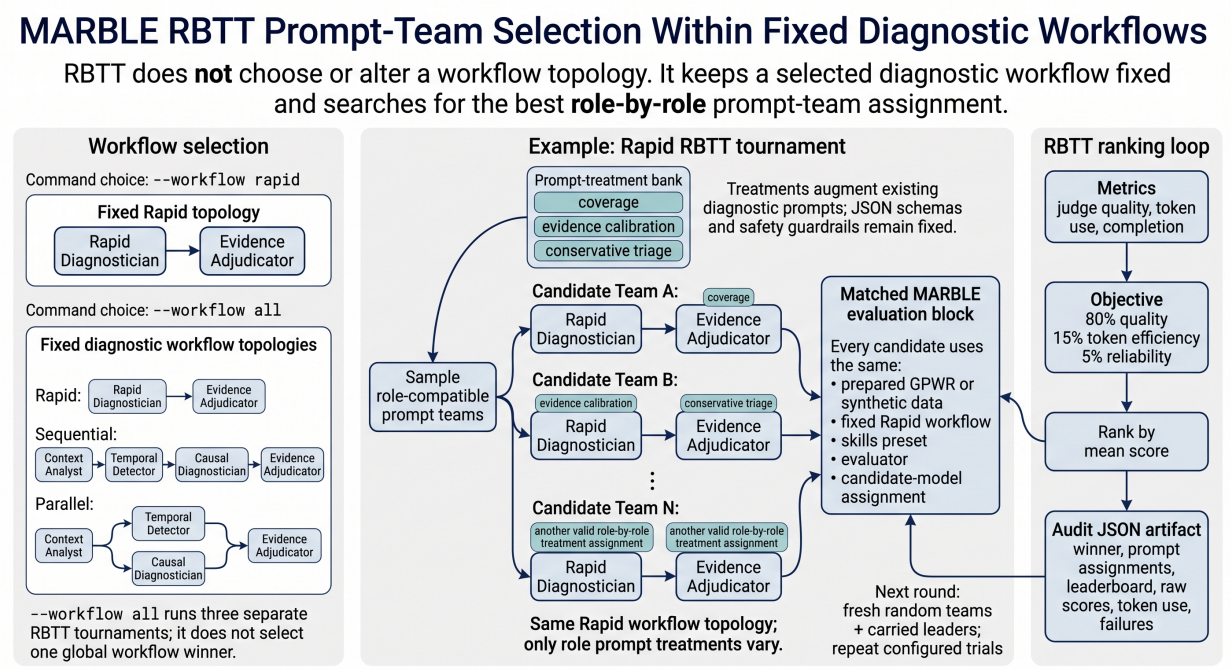


Figure 1. Prompt-team RBTT holds a diagnostic workflow fixed while varying role-specific treatments.

3.2 Static Sequential Randomized Block Tournament

The follow-on program compares model-and-Skill Candidates in the four-role Sequential workflow. Each role has two candidate models and three Skill presets, which creates six Candidates per role, 24 Candidates overall, and 1,296 ordered Teams. The Development profile budgets 40 Team-Block Runs. The Pilot profile budgets 120 Runs and reserves held-out Task Blocks for confirmation. Neither profile attempts exhaustive evaluation of the Team registry.

Initialization freezes the candidate registry, Team registry, prompts, ordered Skill contents, data partitions, evaluator, model resolution, schedules, limits, and execution mode into an Experiment Fingerprint. The benchmark reserves 51 normal trajectories for the Normal Reference and assigns nine other normal trajectories to nine disjoint six-case Task Blocks. Hidden labels stay inside the trusted benchmark boundary. Public payloads use opaque case identifiers and exclude source filenames, fault family names, levels, and incident timing.

3.3 Scoring and Selection

Static Sequential RBTT uses ground truth as its primary measure. Diagnosis Score gives 30 percent credit to anomaly classification, 35 percent to subsystem localization, and 35 percent to fault-mode identification. A fixed Evidence Judge supplies a separate secondary assessment of support, calibration, and auditability. Cost and latency break practical quality ties under the protocol's quality-first order. The protocol does not combine these measures into one authoritative weighted utility.

The scheduler first builds broad, balanced coverage. It balances Candidate appearances by role, avoids duplicate exact Teams, and increases cross-role pair coverage. A later Focus phase evaluates quality leaders and declared challengers on matched blocks. The Pilot freezes a provisional leader before the held-out schedule opens. If the leader fails the confirmation rule, the Experiment records selection not confirmed; it does not substitute a challenger after seeing held-out results.

4.0 Software Architecture

4.1 Diagnostic Workflows

The agent stack separates diagnostic responsibilities into domain-neutral roles. The Context Analyst frames plant state against the normal reference. The Temporal Detector identifies departure from normal behavior. The Causal Diagnostician localizes the subsystem and fault mode. The Evidence Adjudicator reconciles claims, alternatives, uncertainty, and limitations.

The Rapid workflow combines context, temporal, and causal work in one Rapid Diagnostician, followed by the Evidence Adjudicator. The Sequential workflow passes structured outputs through all four roles. The Parallel workflow sends Context Analyst output to the Temporal Detector and Causal Diagnostician concurrently, then reconciles both branches in the Evidence Adjudicator.

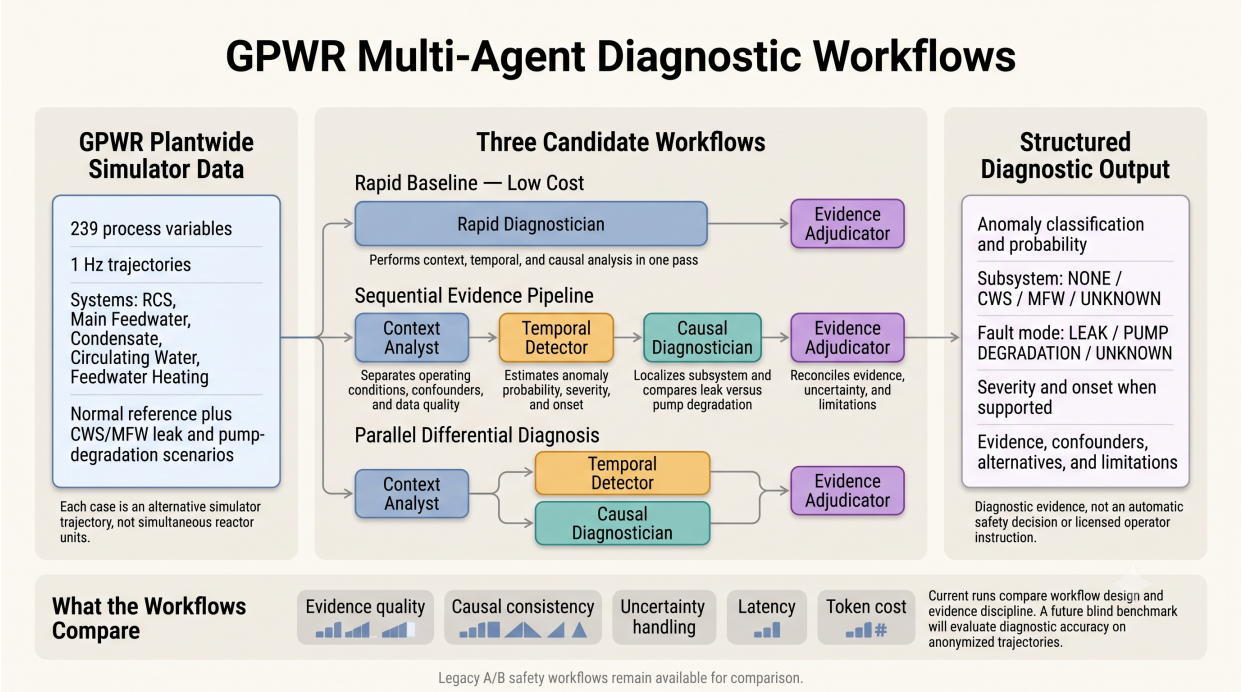


Figure 2. GPWR data enters one of three maintained diagnostic workflows and produces a structured report.

4.2 Programs and Shared Components

Table 3. Program boundaries in the MARBLE repository

Program	Entry point	Question answered
Reactor demo	reactor-demo	How do workflow and whole-team Skill choices affect one diagnosis?
Prompt-team RBTT	marble-rbtt	Which short prompt assignment leads among observed teams in one fixed workflow?
Static Sequential RBTT	rbtt_seq.py	Which observed model-and-Skill Team leads on blinded Task Blocks?

All three programs share the reactor data adapters and agent implementations. The prompt-team engine is domain-free: it samples teams, forms matched blocks, accumulates scores, carries leaders, and ranks means without depending on reactor concepts. The MARBLE adapter prepares GPWR input, invokes the selected workflow, calls the fixed evaluator, and writes the audit artifact.

4.3 Skills and Dependency Injection

GPWR domain knowledge is provided through Markdown-based Skills rather than embedded in agent code. The maintained Skills cover dataset context, system diagnosis, and diagnostic reporting; the all-skills preset adds physics and reporting material. The no-skills, targeted-skills, and all-skills presets control which documents each role receives. Another domain can use the same orchestration with a new knowledge package and data adapter.

The static Sequential executor injects each role's model client, Skill bundle, and mock or live behavior. It rejects a runtime that conflicts with the frozen execution mode, preventing process configuration from changing one Candidate after the Experiment begins.

4.4 Experiment State and Auditability

An append-only JSONL event log is the authoritative Experiment state. The system writes a durable attempt-start event before each provider request and records the terminal outcome separately. Derived registries, schedules, reports, and Task Block projections can be deleted and recreated from the event history. Resume skips completed attempts and stops on an ambiguous paid request until an operator attaches the recovered provider response or marks the attempt failed.

Verification checks the fingerprint, partitions, schedule balance, event hashes, Team traceability, evaluator completeness, held-out isolation, resource records, and credential absence. Simulated and live Experiments cannot share history. Raw GPWR files and Experiment artifacts remain outside version control.

5.0 Results

5.1 Delivered Capability and Verification

The FY2026 project delivered the diagnostic agent stack, Skill injection, the three program entry points, a prompt-team tournament engine, and the complete static Sequential RBTT lifecycle. Supporting material includes the product requirements, normative protocol, operator guide, algorithm background, dataset description, domain model, and two architecture decision records.

A repository verification run on September 15, 2026 selected 187 default tests. All 187 passed in 41.28 seconds; five tests marked for exhaustive profile coverage were deselected by the default configuration. The passing suite covers tournament behavior, benchmark construction, initialization, live and simulated execution boundaries, Pilot logic, release readiness, reporting, recovery, resource policy, selection, compatibility, and reactor workflows.

5.2 Live Prompt Team Tournament

The live study used the Rapid topology, *gemma-4-26b-a4b* for both candidate roles, the targeted-skills preset, and *claude-sonnet-5* as the fixed evaluator. Every candidate received the same six GPWR trajectories. Three rounds produced 32 candidate executions. The tournament directly observed eight of the nine possible prompt teams, made 64 candidate-agent calls and 32 evaluator calls, used about 437,000 tokens, and recorded no execution failures. Teams were ranked by a composite objective combining evaluator-assessed quality, token efficiency, and successful completion.

Table 4. Live Rapid prompt-team leaderboard

Rank	Rapid treatment	Evidence treatment	n	Objective	Judge	Tokens
1	coverage	evidence calibration	2	0.924	10.0	13,644
2	conservative triage	evidence calibration	4	0.913	10.0	13,626
3	conservative triage	conservative triage	2	0.911	9.0	12,189
4	conservative triage	coverage	6	0.907	10.0	13,549
5	evidence calibration	coverage	6	0.881	10.0	13,907
6	coverage	conservative triage	4	0.879	10.0	14,049
7	evidence calibration	conservative triage	2	0.862	9.0	13,007
8	coverage	coverage	6	0.849	9.7	13,974

The leading observed team assigned coverage to the Rapid Diagnostician and evidence calibration to the Evidence Adjudicator. Its mean objective was 0.924 across two observations. The objective range across the eight observed teams was 0.849 to 0.924. Sampling with replacement left the evidence-calibration and evidence-calibration assignment unobserved.

5.3 Observed Patterns

Evaluator quality dominated the objective and often reached its ceiling. Twenty-nine of 32 reports received 10 out of 10. The evaluator consistently rewarded calibrated probabilities, named confounders, credible alternatives, and explicit data limits. Each recorded deduction concerned missing confounders or alternative hypotheses. With quality tied for many reports, token efficiency changed the ordering.

Role-level sample averages differed. Conservative triage had the highest mean objective in the Rapid Diagnostician role at 0.910 across 12 observations. Evidence calibration had the highest mean in the Evidence Adjudicator role at 0.917 across six observations. The repeated coverage assignment produced the lowest exact-team mean at 0.849 and among the longest reports. These differences suggest that the value of a prompt treatment may depend on the agent role in which it is applied. These marginal averages describe the observed sample; they do not isolate a causal contribution from one treatment.

5.4 Limits of the Evidence

The live tournament establishes that the workflow and tournament ran end to end on the PNNL AI Incubator endpoint under the recorded configuration. It also identifies one leading observed prompt assignment under the declared objective. The retained seed, schedule, prompts, raw reports, judge explanations, tokens, and scores make the search auditable.

The study does not establish diagnostic accuracy because the evaluator did not receive ground-truth fault labels. Six trajectories, one candidate model, one Skill preset, and one topology cannot support a generalization claim. The leading team had two observations, the full spread was 0.075, and the study did not test statistical significance. The term leading observed team is therefore the appropriate result label.

The static Sequential lifecycle supplies the design needed for the next evidence stage, including blinded Task Blocks and Diagnosis Score. Its current results are simulated software validation. They confirm the mechanics of registry construction, scheduling, execution, recovery, replay, reporting, and held-out logic; they provide no live benchmark ranking.

6.0 Conclusions and Next Steps

6.1 Conclusions

MARBLE demonstrated a systematic approach for evaluating LLM agent teams through randomized-block tournaments under matched conditions. The software separates workflow execution, domain knowledge, team search, evaluation, and Experiment state. That separation made it possible to run a live prompt-team tournament while keeping the candidate model, data, Skills, evaluator, and topology fixed.

The live result supports two practical findings within its sample. Prompt treatment placement corresponded with different mean objectives by role, and the evidence-quality rubric reached its ceiling too often to serve as the only quality measure. Token use resolved many ties. The outcome justifies ground-truth confirmation rather than a broader performance claim.

6.2 Next Evidence Stage

The immediate next step is a live 40-Run Development Experiment with the static Sequential protocol. It will evaluate diagnostic accuracy against ground-truth labels, testing whether judge-preferred teams also correctly identify anomalies, subsystems, and fault modes on blinded Task Blocks. A later 120-Run Pilot can freeze a provisional leader and evaluate it on held-out blocks. Both activities require an operator-triggered paid run and should retain the current fingerprint, budget, retry, and reporting controls.

6.3 Broader Evaluation

Further work can run separate prompt-team tournaments for the Sequential and Parallel topologies, widen model-and-Skill Candidate choices, and report quality-cost tradeoffs without replacing the protocol's quality-first order. A sponsor-facing pilot should use one defined diagnostic task and preserve the same evidence boundaries. Transfer to another plant or process would require a new data adapter, domain Skills, task labels, and Experiment Fingerprint.

Appendix A: Reproducibility Record

The commands below identify the maintained paths used to reproduce software checks and to start the two tournament programs. Live commands require authorized provider credentials and operator approval. Simulated commands exercise software behavior without creating benchmark evidence.

A.1 Repository verification

```
uv run pytest
```

Observed on September 15, 2026: 187 passed, 5 deselected, 41.28 seconds.

A.2 Prompt team tournament

```
uv run marble-rbtt --workflow rapid --skills targeted_skills --model gemma-4-26b-a4b-project --rounds 3 --teams-per-round 6 --trials 2 --seed 7
```

A.3 Static Sequential simulated lifecycle

```
uv run python rbtt_seq.py run-all --config configs/rbtt_sequential_simulated.yaml --mock
```

A.4 Static Sequential live lifecycle

Use configs/rbtt_sequential_live.yaml after resolving the provider endpoint, credential variable, available models, data directory, prices, and budget. Omit --mock for live execution.

A.5 Evidence labels

- The FY2026 Rapid prompt-team result is live evidence-quality and resource evidence.
- Static Sequential mock results are labeled SIMULATED and validate mechanics only.
- A live run without the independent Evidence Judge is labeled DIAGNOSIS_ONLY.
- Only directly evaluated teams may appear as a Leading Observed Team.
- A Pilot that fails its held-out rule records SELECTION_NOT_CONFIRMED.

Pacific Northwest National Laboratory

902 Battelle Boulevard
P.O. Box 999
Richland, WA 99354

1-888-375-PNNL (7665)

www.pnnl.gov