

STUDY SUMMARY / PREPARED 22 SEPTEMBER 2026

Ten-run repeatability UCMJ evaluation evidence

Two preserved ten-run studies document consistency under recorded test conditions. This readable edition brings together the results, scoring history and evidence references.

UCMJ means Uniform Code of Military Justice, the source material used to generate the 238-question panel.

97.48%

Mean July pipeline accuracy
Adopted scoring basis

236 / 238

Question verdicts stable
Across all ten July runs

10 + 10

Saved runs in two studies
May and July 2026

July: consistency on the adopted scoring basis

The July 25 cohort averaged **97.48%** pipeline accuracy, with a **97.06-97.90%** range. The sample standard deviation was **0.28 percentage points**. The same stored verdict appeared in all ten runs for **236 of 238 question IDs** (**99.16%**).

The comparison covers **4,760 stored observations**: 238 questions, ten runs, and two arms. Mean advantage over the vanilla-RAG baseline was **3.44 percentage points**, ranging from 2.52 to 4.20 points.

May: the original two-host study

All ten May 26 runs recorded **97.48% raw pipeline sweep accuracy**. Five runs were labeled S1 and five S2. The historical adjusted headline ranged from 97.03% to 97.48% under a scoring rule subsequently marked superseded.

How to interpret the result

These observations support repeatability under the tested conditions. Repeated verdicts do not by themselves establish correctness, performance on other workloads, independent certification or accreditation. May raw sweep results and July adopted headline results are separate measures.

Prepared from preserved records and arithmetic comparisons. No models were called and no answers were rescored to create this study edition.

JULY 25, 2026 / ADOPTED FINAL SCORING BASIS

Ten runs, one recorded test panel

All ten saved configuration snapshots are byte-identical, as are the ten sample manifests. Each adopted headline scores all 238 questions, with no questions or connection errors excluded.

Run	Saved run folder	Pipeline %	Baseline %	Advantage (pp)	Archive package ID
1	20260725_143846	97.48	93.70	+3.78	run_f71324ab9c999d681e2c11bb
2	20260725_145307	97.48	93.70	+3.78	run_0d0d2a7419eea261d95c51b4
3	20260725_150728	97.48	94.54	+2.94	run_6adc62a84e681f962a08a2fa
4	20260725_152205	97.06	93.70	+3.36	run_e3eee2f8b9bc87b77987a89e
5	20260725_153644	97.90	93.70	+4.20	run_71f3d0dcb8b8e645ac3f64a7
6	20260725_155127	97.48	94.12	+3.36	run_b162ceb07c73a73ea9059107
7	20260725_160505	97.48	94.96	+2.52	run_8d3dc244c068ef08eb5c42bf
8	20260725_161954	97.06	93.28	+3.78	run_88955f29a694607349924506
9	20260725_163448	97.90	93.70	+4.20	run_94d0114b9fa4adf7e725ce32
10	20260725_164847	97.48	94.96	+2.52	run_edfcfbf1e57b6d53d105c1dd

Recorded configuration and scoring

The model under test and first-pass Doctor are recorded as **Qwen3-30B-A3B-Instruct-2507**. Configuration uses full-panel sampling, temperature 0.1 and max_tokens 200. Adopted headlines state that the deterministic validator and code pre-filter decide results, with LLM judgments advisory.

Corrections and preserved versions

These are the adopted figures after documented refusal-scoring corrections. Run 8's baseline changed from **92.86%** to **93.28%** on August 26; its pipeline figure remained 97.06%. Pre-correction packages remain preserved and are linked in the study JSON.

All ten selected package manifests pass their file-hash checks. Original signature-authority trust was not revalidated.

Answer-text comparison

125 of 238 question IDs have more than one pipeline answer wording across the ten runs. No same-question, identical-text answer has conflicting stored verdicts in this set. These counts agree with the preserved August 26 replay summary.

MAY 26, 2026 / ORIGINAL TWO-HOST STUDY

Stable raw sweep results

The preserved text summary reports two separate A100-80GB servers. Folders identify five S1 and five S2 runs. Host separation is a recorded study claim, not newly independently attested in this edition.

Run	Recorded host	Raw sweep %	Historical adjusted headline %	Archive package ID
1	S1	97.48	97.06	run_60eb77c91a7eb161d6ffc067
2	S1	97.48	97.48	run_97655cb5d25ffd83a6b29900
3	S2	97.48	97.06	run_585ca5ab11dc651c4bc6ca7a
4	S1	97.48	97.05	run_580eac9ac21cde738b4f9119
5	S2	97.48	97.03	run_a2bada2b3ca6dd7596eeb4a6
6	S1	97.48	97.05	run_2c6fbc387d5491ac3d486d84
7	S2	97.48	97.03	run_187022a3f28daaeb0dd6cd75
8	S1	97.48	97.48	run_569375e37e88f8cbd6d64796
9	S2	97.48	97.46	run_3fe0ee21057973bc60729611
10	S2	97.48	97.05	run_5612244816f48e253a8d659d

Scoring basis matters

The historical study explicitly marks its May headline rule as superseded. The zero-spread figure applies to the directly preserved **raw sweep metric**. It does not describe July's headline or show that every generated answer was identical.

Trace the evidence

The archive retains the original May CSV, text summary, HTML/PDF study, configurations and individual results. For exact source IDs, SHA-256 hashes, configurations and correction history, consult:

featured-repeatability.json - July cohort

featured-repeatability-may.json - May cohort

veritrooper-evidence.sqlite - original files and indexed evidence

The package IDs in both tables identify the selected preserved inventories. The database links each inventory to its original artifacts and located copies. Copies and revised packages do not establish additional independent executions.

Source edition

This PDF is a readable edition of repeatability-study.md in the evidence archive dated 22 September 2026. The source Markdown and machine-readable study records remain unchanged. It is an evidence summary, not a certification or accreditation decision.