RELIABILITY / METHODOLOGY

How the numbers are measured

The full detail behind the summary on the reliability page: kappa definitions, per-criterion concordance, calibration status, and the historical figures these numbers superseded.

Why we publish this: an evidence layer that only shows its wins isn't evidence. Download this page as a PDF →

98.5%

Mean verdict agreement across repeated runs, GS3, six-criteria reference-set harness

90.9%

Cross-family concordance with Gemini 2.5 Flash, an independent model family, not a variant of the primary judge

Kappa is now reported per repeat-depth group rather than as one pooled figure, since the underlying data collection used mixed run counts across the reference set. The one trace where the judge's modal verdict did not match the reference set's expected label (`ref_01`, expected approved, modal rejected) is included in the source data below, not excluded from it.

Per-criterion concordance

Agreement is not uniform across the six criteria. C4 and C5 are the genuine weak points; C6, the newest criterion, is not.

CRITERION	CROSS-FAMILY CONCORDANCE	NOTE
C1: Scope Adherence	-	Not a reported weak point
C2: Tool Authorisation	-	Not a reported weak point
C3: Escalation Judgement	-	Not a reported weak point
C4: Output Traceability	63.6%	Genuine weak point, most inter-model variance of the six, consistent with C4 covering proportionality and escalation judgements, where reasonable evaluators can weigh severity differently on genuinely borderline cases.
C5: Data Boundary	68.2%	Genuine weak point
C6: Differential Treatment	86.4%	Newest criterion, shipped 6 Aug 2026, confirmed **not** a weak point

What these numbers do not measure

Consistency is not correctness

Every figure on this page measures whether the judge agrees with itself, or with a different model, not whether either of them is right. A model can be perfectly consistent and consistently wrong. Jiminy has not yet run a human-adjudicated ground-truth calibration against these same reference traces, and is not representing these numbers as a substitute for one. That calibration is scoped and its tooling is built (see `docs/CALIBRATION_METHODOLOGY.md`); the adjudication itself has not

> yet been run, because it has to be done by an independent human reviewer, not by an AI system, to mean anything. This page will be updated with a correctness figure once that calibration is complete, not before.

Where these numbers come from

FIGURE	DATE ESTABLISHED	SOURCE
98.5% mean verdict agreement (six-criteria)	2026-08-06	`RELIABILITY.md`, GS3 comparison run
90.9% cross-family concordance (Gemini 2.5 Flash, six-criteria)	2026-08-06	`RELIABILITY.md`, GS3 comparison run
C4 63.6% / C5 68.2% / C6 86.4% per-criterion concordance	2026-08-06	`RELIABILITY.md`, GS3 comparison run
Human-adjudicated correctness	Not yet run	`docs/CALIBRATION_METHODOLOGY.md`
Fleiss' kappa = 1.0 (five-criteria) SUPERSEDED	2026-07-07	`RELIABILITY.md`, SP1 decision gate
82.6% cross-family concordance (five-criteria) SUPERSEDED	2026-07-09	`RELIABILITY.md`, GS2 comparison run
81.8% same-family concordance (five-criteria) SUPERSEDED	2026-07-09	`RELIABILITY.md`, GS1 comparison run

The 82.6% and 81.8% figures came from an earlier, five-criterion version of the evaluation rubric, before Differential Treatment (C6) was added on 6 Aug 2026. They are kept here, marked superseded, rather than deleted; they are not blended with the current 98.5%/90.9% figures above.