

NyayaVeda AI: legal accuracy benchmark

20 October 2026, release r91: 600 questions (the first 100 of each of six areas, in the seeded order of the same plan the 1 October 1,959-case sample was drawn from), asked of the production pipeline on the Enterprise plan with the answer cache off, and read by two validated readers. Each figure is k of n with its 95% interval, so every number can be checked against its population.

Area	1 Oct 2026 1,959-case sample	20 Oct 2026 (release r91) 600 cases of the same plan
Citations grounded in the sources given (higher is better)	76.0%	87.6% (1964/2242) 95%: 86.0-89.2%
IPC/CrPC to BNS/BNSS transition (higher is better)	76.2%	86.0% (86/100) 95%: 77.9-91.5%
Fabricated citations rejected (higher is better)	99.5%	97.0% (97/100) 95%: 91.5-99.0%
Hindi and Hinglish questions (higher is better)	95.5%	98.0% (98/100) 95%: 93.0-99.4%
Personal data leaked into answers (lower is better)	19.3%*	0.0% (0/95) 95%: 0.0-3.9%
Overruled law pleaded as good law (lower is better)	9.2%	12.0% (9/75) 95%: 6.4-21.3%
Split-bench: which opinion binds (higher is better)	52.6%	47.0% (47/100) 95%: 37.5-56.7%
Overruled / per incuriam recognised (higher is better)	50.3%	52.0% (52/100) 95%: 42.3-61.5%

How to read it. With 100 questions an area, a difference from the 1 October sample is established when the 1 October value lies outside the 95% interval shown. Citation precision counts every checkable citation, so it is the most precise row. * PII leakage of 1 October was measured before the PII-N1 fix (0.3% after it).

What changed since 1 October, each traced to its root cause (the full record is in the task list): a correct answer citing the user's own citation of a judgment it was given is no longer refused as unverified; one citation spelled two ways is counted once; the benchmark process runs as a serving process does (an unpinned run read an unknown load as busy and skipped the step that names the statutory provisions); and the October repairs to recorded treatment (false overrulings removed, displacements recorded) and to the Supreme Court's share of the sources.

Run k14-r91-600-20261020. Rs 3,243.97 for 600 answers (Rs 5.41 an answer); end-to-end latency median 32.0 s, p95 50.4 s; answered / declined by area: bns 99/0; incuriam 77/23; pii 95/5; quorum 64/36; traps 54/46; vernacular 99/1. A decline is the system saying the sources do not support an answer.

Data. Every answer, both readers' grades and the scores are kept beside this page in the production ingest bucket under probe/bench14/k14-r91-600-20261020/ (answers, readings, score.json, score.md). Confidential: NyayaVeda AI.