



Metrics, Experimentation & GTM

How I will know the AI drafts help: measured in shows, on a ₹0 budget, with a small and honest sample

Jain Sanyam Kamlesh, Product Manager & builder Baselines from the three September 2026 shows

1. North Star metric

AI-drafted questions that hosts approved and that were played in a live show, per month.

Why this metric

A draft is only worth something once a host has *trusted* it enough to approve it, and it has been *played* in front of a hall. Counting drafts or batches would reward volume, even volume nobody uses.

User value

Each approved, played question is one the host did not have to write from scratch. Each comes with an explanation that teaches the hall something after the reveal.

NGO value

More fresh, well-pitched shows from the same 1–2 hosts, and more teaching moments. That is the NGO's mission, at ₹0 cost.

Context, the platform's own North Star: **players per show who answer at least one question**. Baseline: 49 of 59 (83%) on 19 Sep and 33 of 37 (89%) on 25 Sep. The AI feature must not lower it.

2. Supporting KPIs

Metric	What it measures	Why it matters	Baseline → target
Host prep time per round product	Minutes from "start writing" to a ready round	The friction we set out to remove	~60 min → ≤ 15 min
Response rate per question product	Answers ÷ phones joined, per question	Engagement must not fall with AI rounds	57% / 78% → ≥ 70%
False-"wrong" rate, typed answers product	Correct answers scored as wrong	Fairness to players without a Hindi keyboard	23% (53/234) → < 5%
Host approval rate AI	Drafts added ÷ drafts generated	Are the drafts worth reviewing?	new → ≥ 70%
Difficulty calibration AI	Player accuracy by AI level	Fixes the age gap (File 01, Q5)	new → easy ≥ 70%, hard ≤ 45%
Checker false-pass rate AI	"Agreed" drafts a scholar finds wrong	Is the safety net real?	new → < 2%

Already instrumented: every added question stores its level, page, verdict, "edited?" flag and model (File 04), and every show stores each answer with its time and score in Firebase. The approval rate needs one more counter (drafts generated per batch), which is in the backlog.

3. AARRR funnel mapping (users = hosts)

Stage	User action	Metric
Acquisition	The host notices " ✨ Draft with AI <i>Beta</i> " on the Content tab, or in the guide or a WhatsApp demo	% of active hosts who open the dialog in a month

Stage	User action	Metric
Activation	Gets a free key, runs a first batch, and adds at least one question	% of hosts who add an AI question in their first session · time to first added question
Retention	Uses AI drafts again for the next show	% of shows with ≥ 1 AI round · batches per host per month
Revenue	Not applicable: a free NGO tool with no fees and no ads	Proxy: volunteer hours saved per month (prep time saved $\times$ rounds) at ₹0 cost
Referral	A host tells another organiser or pathshala, who asks for access	New host accounts requested · partner groups onboarded

4. A/B test design

Constraint, stated honestly. There is one NGO, 1–2 hosts, and halls of 30–50 players. A classic user-split A/B test is not possible. Instead I use a **within-show test**: the same audience plays both kinds of round, and the order swaps between shows.

Element	Plan
Hypothesis	Host-approved AI rounds keep players as engaged as hand-written rounds (response rate no more than 5 pp lower) and cut host prep time per round by at least 50%.
Control (A)	One round written by hand, as today
Variant (B)	One round of the same type and length drafted with AI, then reviewed and approved by the host (the same person), from the same chapter
Target users	Every phone player at the next 3 shows (about 30–50 each), plus the host(s) preparing them. A then B in show 1, B then A in show 2, A then B in show 3.
Primary metric	Response rate per question (answers $\div$ phones joined)
Secondary metrics	Host prep minutes per round (timed) · accuracy by level (calibration) · share of reveals with an explanation read out · hall reactions collected after the show (as in File 01)
Guardrails	Zero doctrinal errors (a scholar checks every AI item before it is played) · false-"wrong" rate on typed answers $< 5\%$ · show length not more than 10% longer · platform North Star (share who answer ≥ 1) not lower
Success criteria	Prep time -50% or better (a large effect, visible even in 3 shows), and the response-rate difference (B – A) has a 95% confidence-interval lower bound above -5 pp. With the guardrails held.
Sample-size logic	Baseline response rate is about 65% (57% and 78%). To detect a 10 pp difference at $\alpha = 0.05$ (two-sided) with 80% power needs about 356 answers per arm . One show gives about 30–50 players $\times$ 10 questions $\approx$ 300–500 answers per arm. Answers from the same player are correlated (design effect ≈ 2), so about 700 per arm $\approx$ 2–3 shows . I plan 3. Proving <i>non-inferiority</i> to 5 pp would need roughly 4 $\times$ more, so the engagement result is directional , and the prep-time result carries the decision.

A second test that needs no players: shadow scoring

Re-score the 234 real answers from 19 Sep twice: once with the answers as they were, once with AI-generated accepted spellings added. **Expected:** most of the 41 correct answers lost to English letters and longer forms are recovered (12 of the 53 were already recovered by the 23 Sep matcher update), **with no wrong answer becoming right**. A false-right count above 0 fails the test. This runs on past data, so no player is ever scored unfairly to prove the point.

5. Dashboard plan (5-3-1)

Layer	Metrics shown	Who looks	How often
5 · Operate	Response rate per question · AI approval rate · host edit rate · false-"wrong" rate · AI latency and fallback rate	Me (PM) and the hosts	After every show
3 · Steer	Players per show who answer ≥ 1 · shows run per month · volunteer prep hours saved	NGO committee	Monthly
1 · North Star	AI-drafted questions approved and played this month	Everyone	Monthly

Data already exists in Firebase (show answers) and in each game (the "ai" block on every AI-written question). The first version can be the host panel's existing Dashboard page plus a monthly export. No paid analytics tool is needed.

6. Rollout plan

Phase 0 · Internal ✓ done

14–15 Oct · real book, 6 runs, failures fixed, fault injection, shipped as Beta (hosts only)

Phase 1 · One host, one round

Next 2–3 shows: one AI round next to hand-written ones; a scholar checks every AI item

Phase 2 · Every NGO host

The other host joins; up to half the rounds; monthly scholar spot-check

Phase 3 · Partner groups

Pathshalas and youth groups, each with their own free key; any share of rounds

Phases advance only if the North Star grows and every guardrail holds. "Percentage of users" makes little sense with one or two hosts, so rollout is staged by host and by share of rounds.

Kill-switch criteria

- Any AI-written question confirmed wrong **after it was played**: pause immediately
- Checker false-pass > 2% in a scholar audit
- Approval rate < 40% over 3 consecutive batches (the drafts are wasting hosts' time)
- Any cost appearing anywhere** (the NGO's hard rule)

How fast: remove the button with the normal deploy, about 2 minutes. Questions already added stay usable. A remote on/off flag is next in the backlog.

Red-teaming before Phase 2

- Index, preface and blank pages (must say "nothing to ask")
- Pure Prakrit or Sanskrit verse pages, and blurry scans
- Number and list facts (14 गुणस्थान, 7 समुद्रघात)
- Pages where traditions differ
- "Ignore your instructions" text planted on a page (prompt injection)
- Wrong language selected, maximum page range, copyright refusal

7. Go-to-market

Target audience	First the NGO's own hosts (1–2 today), then Jain pathshala teachers and youth groups who run quizzes at events
Positioning	For Jain event hosts who spend an hour writing questions from granths, ✨ Draft with AI turns a chapter into checked, explained, well-pitched questions in about a minute. Nothing reaches the projector until you approve it.
Key message	"पुस्तक दीजिए, प्रश्न पाइए — जाँचें आप।" (Give the book, get the questions, and you do the checking.)
Channels	In-product Beta badge · the Games Guide (section written) · live demo at the next volunteer meeting · organisers' WhatsApp groups · a 2-minute Hindi screen recording

Launch checklist	✓ Free key in a no-billing project · ✓ guide updated · ✓ fault-injection test passed · ☐ scholar named for spot-checks · ☐ red-team pass · ☐ first AI round mixed with hand-written ones · ☐ backup downloaded before every show
-------------------------	--

8. Cost note

₹0, by design. Each batch is **2 requests** (draft and blind check). On the free tier the newest model allows about 20 requests a day, and the feature can fall back across up to 8 free models, so a host preparing even several rounds a day stays inside the free allowance. The Firebase project is on a pay-as-you-go plan but is **never** used for AI calls. The key belongs to a separate Google project with no billing, and paid models refusing that key in testing confirmed it.

Value against cost: about 45 volunteer minutes saved per round, for nothing. **What I watch:** quota errors, how often a lighter model writes the batch, and any change to Google's free-tier terms. If the free tier ended, the decision would go to the NGO committee: fewer pages per batch, or a small donor-funded key. It would never go to the billed Firebase project, and the blind check would never be dropped to save a call.

9. Post-launch feedback plan

Channel	What is collected	How it improves the feature
Review-screen behaviour	Approved, edited and unticked per draft, by level, round type and model (stored automatically)	A high edit rate on one kind of question becomes a new prompt rule. The "नए is never easy" rule came from exactly this.
Checker disagreements	Drafts the blind check rejected, and whether the host agreed	They become regression test cases for prompt changes
Show data	Response rate and accuracy per question and level	Calibration: re-tune the difficulty rubric if "easy" questions score below 70%
Host debrief	Three questions after each show: time saved, what was wrong, what was missing	Backlog priorities
Hall feedback and scholar audit	Comments like those in File 01, plus a monthly check of played AI items	Kill switch, and trust: the most important signal