



Responsible AI & Risk Plan

Religious teaching in front of a live hall: a confident wrong "fact" is a harm, not a typo

Jain Sanyam Kamlesh, Product Manager & builder Covers ✨ Draft with AI (Beta), shipped 14 Oct 2026

Design stance. The AI **drafts**, a second AI **tests** it blind, and the host **decides**. Nothing reaches a show without a human tick. Several guardrails below exist because I watched the first version fail in testing. Those failures are marked **Observed**.

1. Possible AI failure cases

#	Failure	How it shows up in this feature	Seen?
1	Plausible nonsense passes a "check"	14 Oct, two-clue style on a lighter model: "पुद्गल कर्मों का कर्ता तथा संसार के जीवों का सातवां नरक?" → व्यवहार. The two clues do not lead to one answer. The first checker was shown the answer and asked to verify it, and it marked this "found in the book".	Observed
2	Doctrinal hallucination	A wrong count or attribution, for example the number of समुद्घात or who did a deed in a Purana story, is presented with full confidence and then taught to a hall of 30–50 people on the projector.	Risk
3	Difficulty mislabelled	Questions that hinge on निश्चय नय versus व्यवहार नय came back as "easy". Elders and children get lost, which is the age-gap friction from File 01 (Q5).	Observed
4	Answer-position bias	In the first batch the correct option was B in 7 of 10 questions. A regular audience learns that within one evening.	Observed
5	A "proof" quote that is not really in the book	Draft #5 on 15 Oct quoted "समुद्घातादशायां देह प्रमाणं हि रहता है", a garbled mix of Sanskrit and Hindi from reading the scan. It looks like proof, but it isn't a real line.	Observed
6	An over-broad accepted answer	If the AI added भारत as a variant of भरत, a wrong answer would score. The opposite risk: a missing variant keeps marking "Anjana" wrong.	Risk
7	AI unavailable on show day	On 14–15 Oct, four newer models answered 503 "high demand". One request hung for over 3 minutes. The newest model's free quota (20 requests a day) ran out during testing.	Observed
8	Tradition bias	The model mixes in the interpretation of another tradition or sect, or frames a comparison that the NGO would not teach.	Risk

2. Responsible AI risks

Privacy

Book pages go to Google. On the free tier Google may use them to improve its products. The real privacy risk is a host uploading something private, such as an unpublished manuscript or members' notes, or player data being sent by mistake.

Bias and fairness

Two kinds. **Doctrinal**: one tradition's reading presented as the only truth. **Player**: answer-checking that favours people with Hindi keyboards. 19 Sep showed 41 correct answers lost to English letters or longer forms.

Hallucination

The model can write fluent Hindi about things the page does not say. Worse, it can produce a fake "quote". The audience trusts the projector, so one error is taught to everyone at once.

Safety

The harm is not physical. It is **reputational and spiritual**: teaching a wrong tenet at a religious venue in the NGO's name, or a disrespectful phrasing about a tirthankar or another tradition.

User trust

If one wrong AI question is caught on stage, hosts stop trusting the tool. Players may start doubting the human-written questions as well.

Over-dependence

A host in a hurry may tick all ten drafts because "the checker agreed". That turns a safety net into a rubber stamp, and the checker is an AI too, so its mistakes can line up with the drafter's.

3. Guardrails and mitigations

Risk	Possible impact	Mitigation / guardrail (as built)
Hallucinated or wrong fact	Error taught to the hall	Pages-only rule in the prompt. Every draft must give a PDF page and a quote of ≤ 10 words . The blind check must arrive at the same answer independently. 100% human approval ; nothing is auto-added.
A checker that just agrees (case 1)	False sense of safety	Replaced the "verify this answer" check with a blind test . The checker never sees the answer. The code compares its answer using the phones' own matcher, so spelling differences don't count as disagreement but substance does. Disagreed drafts arrive unticked and show what the checker answered. In the two-clue style each clue is answered separately .
Position bias (case 4)	Guessable answers	Options are shuffled in code , not left to the model. The 50:50 pair is picked from the wrong options.
Wrong difficulty (case 3)	Bored or lost groups	Explicit rubric pitched for a family audience ("a question that depends on the नय distinction is never easy"). The level is editable on the review screen. Calibration is measured from player accuracy by level (File 05).
Fake quote (case 5)	False proof	The page number is the real anchor, and the host is told to check the page, not trust the quote. Next step: show a thumbnail of the cited page in the review screen.
Bad answer variants (case 6)	Wrong answers score, or right ones don't	The prompt allows only correct spellings, never broader answers. Variants are shown as an editable list, one per line, before adding. Typed answers still pass through the site's matcher, which tolerates small typos.
Weaker model on busy days	Lower-quality drafts	Model order prefers the full models. The review screen warns: <i>"Google's main model was busy, so a lighter one wrote these. Read them with extra care."</i>
Unavailable (case 7)	No questions on show day	8-model fallback, 2 passes, 120 s time-out per call, and a plain message. The manual paths (+ <i>New item</i> , <i>Add many at once</i>) never depend on AI.
Privacy	Private text or player data exposed	Only the chosen pages (≤ 20) leave the browser, and never any player data . The key is stored only in the host's browser. The screen says: <i>"The chosen pages are sent to Google Gemini. Player details never are."</i>
Copyright	Misuse of a book	Non-commercial teaching only. Quotes are capped at 10 words, and if the model refuses to quote it retries with no quotes. Pages are not stored by us.

Risk	Possible impact	Mitigation / guardrail (as built)
Cost	A surprise bill	Free models only. The key must come from a new Google project with no billing , never the Blaze Firebase project; the dialog and the guide both say so. Paid models refused the key in testing.
Tradition bias (case 8)	Disrespect, mixed teaching	The prompt says: "Follow the book's own view. No comparison with or criticism of other traditions." The source is chosen by the host. Next step: a monthly scholar spot-check.

4. Fallback behaviour

When the AI...	What the host experiences
is unsure	The checker could answer but not from these pages, or one of two clues matched. The draft shows Check this , stays ticked, and gives the checker's answer and reason.
disagrees with itself	The blind answer differs from the draft. It shows Checker disagreed — left out , is unticked , and the host can still edit and tick it.
is unavailable (busy, quota, retired, hung)	A spinner reads "If Google is busy it tries another model...". After the fallback chain, a red banner says "Google's free AI is very busy... You can still add questions by hand or with 'Add many at once'." The settings and book stay loaded. See File 04, Figure 4.6.
checker fails, drafter worked	Drafts still arrive, all marked <i>Check this</i> . The check is a safety net, not a gate that blocks the host.
returns unusable output	The schema forces the shape. The code drops drafts with missing options or answers, and duplicates of existing questions. If nothing usable is left: "The AI found nothing it could ask from these pages. Try other pages or a different focus."
is asked something out of scope	Round types it can't serve (buzzer ordering, photo pairs) get a polite explanation of which rounds are supported. Cover, index and blank pages are skipped by instruction.
refuses to quote (copyright filter)	Silent retry without quotes. The page number still lets the host check.

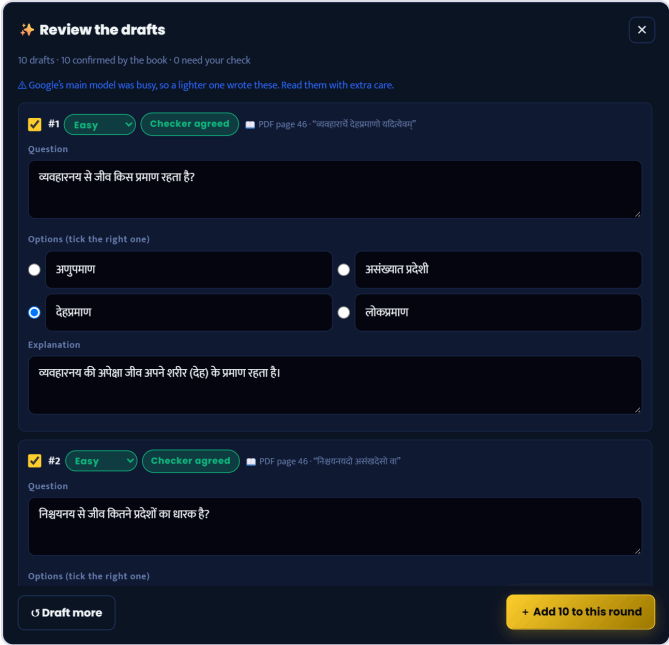
5. Human-in-the-loop plan

- **Every item is approved by the host.** The review screen is the only path into a round. Drafts the checker disagreed with start unticked, and the host sees the checker's answer.
- **Edits are recorded.** Each added question stores whether the host changed it, plus its page, level, verdict and model. A high edit rate on one kind of question tells me which prompt rule to fix next.
- **Live override.** On stage the host controls the reveal and can award points by hand, so a disputed answer can be corrected in the room.

Escalation for sensitive cases

- **Doctrinal doubt** about a draft: do not use it, and ask the NGO's pandit or scholar before the show.
- **Monthly spot-check** (next): a scholar reviews a sample of AI-written questions that were played. Any confirmed error pauses the feature (File 05 kill switch).
- **A player challenges an answer** during a show: the host checks the stored page reference after the show and corrects the question in the Studio.

6. AI transparency



Transparency in the review screen. Beta badge, verdict pills, page and quote, editable level, and the lighter-model warning.

- **Labels:** " ✨ Draft with AI" with a **Beta** badge, and the line "AI reads the pages you choose and drafts questions. **You approve every one.**"
- **Confidence:** three verdict pills from the blind check, the checker's own answer when it disagrees, and a warning when a lighter model wrote the batch.
- **Source:** PDF page and a short quote on every draft.
- **Data notice:** "The chosen pages are sent to Google Gemini. Player details never are."
- **Flagging:** the host unticks or edits. Players raise a doubt with the host in the room. A "report this question" button on phones is in the backlog.
- **On the projector** questions are not tagged "AI". Once a host has checked and approved one it is the host's question, the same as one taken from a reference book. That is the reason the human gate is mandatory, not optional.

7. AI evaluation metrics

Metric	What it measures and why it matters here	Target
Errors that reach a show	AI-written questions later confirmed wrong by a scholar or in the room. This is the harm that matters.	0 (any one pauses the feature)
Checker false-pass rate	Share of "Checker agreed" drafts that a scholar spot-check finds wrong. Tells me whether the blind check really works.	< 2% of audited items
Host approval rate	Drafts added ÷ drafts generated. Shows whether the drafts are worth reviewing.	≥ 70%
Host edit rate	Added drafts the host changed. A proxy for quality, broken down by round type and level.	Falling over time
Difficulty calibration	Player accuracy by AI level in live shows. Easy should be above medium, which should be above hard.	Easy ≥ 70%, hard ≤ 45%
False-"wrong" rate (typed answers)	Correct answers marked wrong. Baseline 19 Sep: 53 of 234 responses (23%).	< 5%
Latency P95	Time from "Draft" to the review screen. Hosts prepare under time pressure.	≤ 150 s
Fallback / failure rate	Batches that needed a lighter model, or failed. Shows how much the free tier can be relied on.	Failure < 5% of attempts
Cost per batch	Must stay free (the NGO's hard rule).	₹0