Kashi Measurement-Science Research Memo

Project-useful summary of the measurement-science research result, tailored to the current Kashi deck and concept note.

Scope	Primary use	Anchor docs
What the research supports, what it does not support, and what Kashi should change now	**Deck revision, product-boundary decisions, pilot design, and judge/investor defense**	**Kashi progress deck (2026-04-21) + meeting governance concept note**

Bottom line. Kashi becomes much more credible when it is framed as a system that estimates repeated interaction asymmetry under uncertainty, within comparable meeting contexts, for review support. It becomes much more fragile when it sounds like a detector of harm, harassment, or abusive intent.

Prepared: 2026-04-21 | This memo is intentionally decision-oriented rather than academic in tone. The goal is to help the team tighten claims, resolve contradictions, and design a defensible pilot.

Executive summary

- The research result is not that Kashi can "detect harm." The stronger and defensible result is narrower: the current concept is best positioned to estimate repeated interaction asymmetry from meeting data under uncertainty, and to package those estimates as review-support signals rather than findings.
- The current Kashi deck already contains several strong moves: it rejects harassment classification, refuses automatic action, refuses performance/HR decision use, uses review-worthy events instead of verdict language, and rejects a company-wide "relationship health" bar. Those choices are aligned with measurement-science logic, not just governance optics.
- The deck still has material weaknesses. The biggest is an internal contradiction: Kashi claims "metadata only / no content," but some named detectors already require semantic processing or transcript interpretation (for example substantive response, topic-credit recovery, or position shift logic). That contradiction should be resolved immediately. Either Kashi becomes truly structural-only in MVP, or it becomes honest about being a constrained hybrid system.
- The strongest research-backed upgrade is a clearer epistemology story: Kashi should say that it estimates interaction asymmetry relative to appropriate baselines, with abstention when evidence is weak. That means moving away from language such as "the pattern is the harm" and toward language such as "the pattern may constitute evidence consistent with uneven conversational treatment."
- The most important technical addition is not another detector. It is a measurement layer: baseline stack, evidence grades, abstention rules, input-quality gating, and a validation plan that tests real-world robustness instead of only planted synthetic scenarios.
- For the project right now, the actionable outcome is clear: narrow the claim, clean up the content contradiction, operationalize confidence, add an explicit abstention policy, and present the pilot as proof of mechanism plus validation roadmap rather than proof of full validity.

Decision matrix: the answers Kashi should give

Question	Best answer for Kashi	Why this is the strongest position
What do we measure?	Repeated interaction asymmetry under uncertainty, not harm itself.	This matches what transcripts and meeting structure can plausibly support; it avoids pretending to measure intent, legality, or psychological state.
Compared to what baseline?	Self-history first, then within-meeting peer comparison, meeting-type baseline, role baseline, and dyad baseline.	There is no credible single universal norm for "healthy meetings." Context is part of the measurement problem.
When is a signal credible?	Only after repeated comparable exposure, with evidence grade and uncertainty attached.	Meeting count alone is not enough; sparse observations should not become person-level claims.
What should the UI show?	Review-support signals, evidence grade, traceable event references, explicit caveats, and abstention where needed.	A naked score invites overconfidence and misuse.
What should Kashi refuse to claim?	Harassment detection, intent detection, illegality, future-behavior prediction, and fully comparable cross-team truth scores.	Those claims are not supported by the evidence available from meeting transcripts alone.
What is the pilot proving?	Proof of mechanism and initial usability of the governance model; not universal validity.	The current evidence base is too small and too synthetic to support stronger claims.

Use this matrix in the pitch. If a judge asks "what exactly are you measuring?", "compared to what?", or "how many meetings until this means anything?", this table gives the disciplined answers. Anything more aggressive than this will make the project easier to attack.

1. What the research actually supports

1.1 Construct validity: the strongest credible claim is narrower than the current rhetoric

The foundational measurement point is simple: a system is not "valid" in the abstract. Its outputs are only as defensible as the intended interpretation and intended use. The 2014 Standards for Educational and Psychological Testing explicitly frame validity around the justification of score interpretation and use, not around a free-floating notion that a score is simply "true." That matters directly for Kashi because the same detector output could be used responsibly for private review support or irresponsibly as pseudo-disciplinary evidence.

Applied to Kashi, this means the product should not claim that it detects harm, harassment, abusive managers, or hostile intent. Meeting transcripts and turn structure do not directly measure those constructs. What they can more plausibly support is a narrower construct: whether repeated meetings show unusually asymmetric conversational treatment patterns relative to an appropriate reference class.

This is why the best conceptual formula for Kashi is: interaction asymmetry risk = an uncertainty-bounded estimate that one participant or subgroup receives systematically different conversational treatment than relevant peers across comparable meetings. That framing keeps the product inside a zone where explanation, challenge, and abstention are still possible.

Key grounding: AERA / APA / NCME, Standards for Educational and Psychological Testing (2014); NIST AI RMF 1.0 (2023).

Practical implication for Kashi. Do not say "the pattern is the harm." Say "the pattern may provide evidence consistent with uneven conversational treatment and may justify human review."

1.2 Baselines and comparability: there is no single normal meeting

A core weakness in many workplace-measurement products is the fantasy of a universal baseline. Kashi is already ahead of that problem because the deck emphasizes per-speaker calibration to a 90-day rolling baseline rather than team average. That is one of the strongest parts of the current design because it directly reduces several obvious false-positive modes such as introversion, second-language participation, or facilitator-heavy meetings.

However, self-baseline alone is not enough. The same person may behave very differently in a sprint review, incident call, governance committee, 1:1 coaching session, or all-hands Q&A. Cross-cultural and virtual-meeting research also shows that participation form and meeting expectations vary materially across groups and contexts. If Kashi treats all meetings as interchangeable, the signal will be contaminated by meeting genre, role structure, language environment, and local norms rather than by asymmetry alone.

The correct answer is a baseline stack. Kashi should evaluate people against at least five references: self-history, within-meeting peers, meeting-type baseline, role baseline, and dyad baseline. Self-history catches personal change; within-meeting peers catch selective treatment inside the session; meeting-type and role baselines reduce false interpretation; dyad baseline catches repeated directional targeting.

Key grounding: NIST AI RMF 1.0 (context and fit-for-purpose); Köhler & Gölz, "Meetings Across Cultures" in The Cambridge Handbook of Meeting Science (2015); Kreamer et al., "Virtual Voices" (2024).

Baseline layer	What it answers	Why Kashi needs it
Self-history	How is this person being treated versus their own prior comparable meetings?	Prevents introversion or naturally low floor time from being misread as suppression.
Within-meeting peers	Is one participant getting a disproportionate burden inside this session?	This is the fastest route to detecting selective treatment in context.
Meeting-type baseline	Is the current interaction normal for this type of meeting?	Incident calls, retrospectives, and 1:1s have different speaking norms.
Role baseline	How do comparable roles usually behave here?	Facilitators, PMs, juniors, and guest SMEs are not behaviorally interchangeable.
Dyad baseline	Does A behave differently toward B than toward others over time?	This is the cleanest route to repeated directional asymmetry.

1.3 Evidence accumulation and uncertainty: credibility depends on exposure, not just meeting count

A likely judge question is “how many meetings before the signal is credible?” The correct response is not a magic number. Measurement theory treats repeated behavioral observation as noisy because multiple sources of error are operating at once: occasion, context, speaker mix, meeting purpose, audio quality, and more. Generalizability theory exists precisely to reason about the dependability of measurements when error comes from more than one source.

For Kashi, that means the right variable is not simply meeting count. The real question is how much relevant exposure exists. Three highly interactive comparable meetings may be more informative than ten meetings in which the person barely speaks. A person with repeated dyadic exchanges yields stronger evidence than a passive attendee. A one-off all-hands should not be allowed to dominate a person-level view.

This directly implies three product requirements: first, Kashi should show evidence grade or uncertainty, not just raw risk score; second, early sparse observations should be pulled toward conservative interpretations; third, the system should abstain when there is not enough comparable exposure. NIST’s recent evaluation work reinforces this logic by emphasizing uncertainty estimates and careful assumptions rather than naked benchmark numbers.

Key grounding: Shavelson, Webb & Rowley, “Generalizability Theory” (1989/1990); NIST AI 800-2 ipd (2026); NIST AI 800-3 (2026).

Low-quality answer Kashi should avoid	Better answer Kashi should give	Why
“After 5 meetings we know.”	“Credibility depends on repeated comparable exposure, interaction opportunity, and input quality.”	This avoids fake precision.
Single composite score only	Composite signal + evidence grade + reason codes	Users need to know how strong the basis is.
Always output something	Abstain when evidence is insufficient or confounded	Abstention is a strength, not a weakness.

1.4 Input-layer reliability: Kashi may partly measure who the speech system fails on

This is one of the most important and underweighted findings. Kashi starts from transcripts, speaker attribution, timestamps, and in some detector formulations, limited semantic interpretation. That means its downstream signal is only as fair as the input layer is reliable. If the system mishears, mis-attributes, or fails on overlaps, then the measurement can become distorted before any detector logic even runs.

The risk is not theoretical. Published work has found substantial disparities in automated speech recognition performance across speaker groups. The widely cited PNAS paper by Koenecke et al. found large racial disparities in commercial ASR, with average word error rates around 0.35 for Black speakers versus 0.19 for white speakers across five systems. More recent work on disfluent speech also found statistically significant bias against stuttered/disfluent speech with syntactic and semantic degradation in transcripts. Even if Kashi’s immediate deployment context is different, the measurement lesson is the same: transcript quality and diarization quality are part of the construct-validity problem, not a side note.

For Kashi, that means confidence must include input quality. The system should degrade gracefully under poor transcript confidence, overlap-heavy segments, low diarization confidence, multilingual confusion, or strong L2 conditions. Some outputs should be blocked entirely when the input basis is too weak.

Key grounding: Koenecke et al., PNAS (2020); Mujtaba et al., NAACL / ACL Anthology (2024).

Kashi should add immediately	Why it matters
Transcript-confidence gating	Prevents low-quality transcripts from quietly becoming pseudo-evidence.
Speaker-diarization confidence gating	Directional interruption metrics are fragile if speaker attribution is wrong.
Overlap-quality flag	Interruption detectors become unstable in heavy crosstalk or poor audio conditions.
L2 / multilingual caveat surface	Participation and latency are partly language-dependent.
Pilot subgroup error audit	Without this, fairness claims are unsupported.

1.5 Goodhart pressure and post-deployment monitoring: metric improvement is not the same as cultural improvement

Kashi is already smart to refuse the company-wide “relationship health” bar. That refusal is not just ethical; it is scientifically wise. Once a visible number becomes the target, users optimize the number rather than the reality. Goodhart’s Law is directly relevant here: if managers know exactly what is being measured, they can learn to look clean while preserving the underlying power dynamic.

This means Kashi cannot treat improvements in interruption rate, speaking-share parity, or directive-density measures as sufficient proof of healthier behavior. Those outputs are useful, but they are gameable. Pressure can move into side channels, follow-up messages, agenda control, selective exclusion, or performative politeness on recorded calls.

The correct response is post-deployment monitoring. NIST’s 2026 monitoring report is especially relevant here: it emphasizes that pre-deployment testing is not enough and that monitoring methods, field studies, and incident-tracking remain underdeveloped but necessary. In Kashi terms, that means pilot success must include not only dashboard movement but also dispute rates, reviewer audits, evidence of issue displacement, and whether people actually trust the system enough to use it.

Key grounding: CNA, Goodhart’s Law (2022); NIST AI 800-4 (2026).

Critical read for the team. A manager mirror can still be gamed even if the company-wide health bar is refused. Refusing one bad metric does not remove the need for anti-gaming design and post-deployment monitoring.

1.6 Validation logic: the pilot is currently proof of mechanism, not proof of general validity

The current deck cites seed scenarios, deterministic reruns, and zero false positives on a healthy control. That is fine for a hackathon proof-of-concept. But scientifically, it only shows that Kashi can detect patterns that were intentionally planted and that it does not trip on one authored healthy case. It does not yet establish robustness, subgroup fairness, confound resistance, or real-world usability of the outputs.

A stronger validation story should be argument-based rather than benchmark-bragging. Kashi does not currently have a canonical public benchmark for “workplace power asymmetry in meeting transcripts,” so the

right path is to build evidence across multiple validity dimensions: construct validity, robustness to input error, convergent validity across detectors, discriminant validity against normal disagreement, stability across time, contestability by users, and consequence monitoring after deployment.

The hardest but most important truth is that the pilot should not claim “we know we are right.” It should claim “we know the system can produce traceable structural signals; the next step is disciplined validation of when those signals are or are not trustworthy.” That is a more serious and more defensible stance.

2. Kashi-specific assessment against the current deck

2.1 What the current deck already gets right

Kashi already rejects some of the worst possible claims. The deck says it is not a harassment classifier, not a legal/intent detector, not an employee-monitoring tool, and not a platform for HR decisions. It uses “review-worthy event” language rather than “problematic statement,” and it explicitly refuses company-wide relationship scoring. These are strong positions and should be preserved.

The move to per-speaker baseline calibration is genuinely important. It is one of the project’s most defensible features because it directly addresses obvious confounds that sink naive engagement or productivity tools.

The governance posture is also unusually disciplined for a hackathon-stage concept. Role-based access, no default transcript browsing, audit trails, and a refusal to promise automatic action are all consistent with the research conclusion that measurement should stay assistive rather than adjudicative.

Keep as-is or keep with only minor wording changes	Why
“Review-worthy event” instead of moral or legal labels	This is the right construct boundary for an uncertain system.
No automatic HR or disciplinary action	This protects the use-boundary and prevents validity creep.
Per-speaker baseline calibration	This is one of the strongest anti-false-positive design choices in the deck.
Refusal of the company-wide relationship-health bar	This is both governance-smart and measurement-smart.
Role-based visibility and audit trail	Needed to keep the system from collapsing into surveillance.

2.2 Where the current deck is still vulnerable

The most serious vulnerability is the rhetoric drift between careful and strong claims. Some sections correctly present Kashi as a support layer for human review; other lines slide toward language that implies direct detection of harm or harmful team dynamics. That drift is dangerous because judges will always quote the strongest-sounding line back at you.

The second vulnerability is the content contradiction. The deck repeatedly says “metadata only,” “never transcribe for analysis,” “none read meeting content,” and “Kashi never reads message content.” But several detectors as currently described require some form of transcript interpretation or semantic comparison. The project must resolve this instead of trying to rhetorically outrun it.

The third vulnerability is the under-developed treatment of uncertainty. The score formula contains “confidence,” but confidence is not yet operationalized. Without a visible evidence model, confidence looks like a black-box stabilizer rather than a measurement discipline.

The fourth vulnerability is evaluation overstatement. The demo evidence is enough to show mechanism, not enough to claim that the detectors already generalize cleanly to real organizations.

Current deck move	Why it is vulnerable	Risk if left unfixed	Recommended correction
“The pattern is the harm.”	It collapses behavioral signal into moral conclusion.	Overclaim; easy attack on construct validity.	Change to “the pattern may reveal evidence consistent with uneven conversational

			treatment.”
“Detecting harmful team dynamics earlier.”	Implies stronger semantic certainty than the evidence supports.	Sounds like a harm detector, not a review-support system.	Change to “surfacing repeated interaction asymmetries earlier.”
“Metadata only / never reads content.”	Conflicts with detectors that infer substantive response, topic-credit, or agreement shift.	Internal inconsistency; credibility loss.	Either remove semantic detectors from MVP or explicitly split structural and constrained semantic detectors.
Confidence inside composite score only	Confidence is not defined for users or judges.	Looks arbitrary.	Add evidence grade, reason codes, and abstention policy.
3 harmful seeds + 1 control as proof	This is mechanism testing, not general validity.	Evaluation overclaim.	Describe as proof of mechanism plus planned validation roadmap.

2.3 The key strategic choice: structural-only MVP or honest hybrid MVP

The project should make this decision explicitly. Right now, the deck wants the legal/ethical benefits of “metadata only” while also enjoying some of the detection power that comes from transcript-level semantic cues. That is not stable. The team needs to pick an architecture and talk about it honestly.

Option A is structural-only MVP. In that version, Kashi limits itself to speaking share, overlap/interruption, latency, chilling-delta, dyadic concentration, and similar timing/graph signals. It gives up topic-credit recovery, substantive-response logic, and position-shift inference unless those can be redefined without semantic content. The benefit is a cleaner legal and epistemic story. The cost is lower sensitivity in subtle cases.

Option B is honest hybrid MVP. In that version, Kashi says that core live detection is rule-based and explainable, with a small set of constrained semantic detectors operating on transcript text but not exposed to employer browsing as content. The benefit is better signal coverage. The cost is that the privacy/governance story becomes more nuanced and must be defended carefully.

For a hackathon-stage deck, structural-only is cleaner. For a real product, honest hybrid may eventually be stronger. But the current mixed rhetoric should be removed.

3. Required measurement model for Kashi

If the team wants a usable framework rather than only critique, this is the operational model to build toward. It is narrow enough to be defensible and concrete enough to guide product and pilot decisions.

Measurement component	What Kashi should do
Target construct	Estimate repeated interaction asymmetry under uncertainty within comparable meeting contexts.
Primary output type	Review-support signal, not finding; event objects and pattern summaries, not moral labels.
Baseline stack	Self-history + within-meeting + meeting-type + role + dyad.
Evidence representation	Composite indicator plus evidence grade, uncertainty notes, and reason codes.
Confidence inputs	Transcript quality, diarization quality, overlap quality, comparable exposure, detector agreement, user-marked confounds.
Abstention policy	Suppress or down-rank output when evidence is thin, confounded, or input quality is poor.
Validation approach	Mechanism testing first, then human-review alignment, subgroup audit, robustness testing, and post-deployment monitoring.
Use boundary	No HR performance use, no automated action, no claim of intent/illegality/harassment detection.

3.1 Evidence grades Kashi should expose in-product

A simple evidence-grade ladder will do more for credibility than another abstract score term. Users need to know not only that a pattern fired, but how much weight they should put on it.

Recommended ladder:

- Insufficient evidence — output suppressed or only shown as background telemetry; not enough comparable exposure, or input quality too low.
- Weak pattern — one or two detectors suggest asymmetry, but exposure is limited or confounds are strong.
- Emerging pattern — repeated signal across comparable meetings, but still unstable or not yet broad enough for strong interpretation.
- Stable pattern — repeated signal across comparable meetings with adequate input quality and low unresolved confounding.

3.2 Minimum abstention rules

The system should not be forced to always say something. That would be bad measurement practice. Minimum abstention rules should include at least the following:

- No person-level interpretation when the comparable exposure window is too thin.
- No directional asymmetry output when speaker attribution quality falls below threshold.
- No interruption or chilling interpretation from overlap-heavy audio with degraded segmentation.
- Strong caveat or abstention in multilingual / L2-heavy sessions until pilot data supports stable interpretation.

- Down-rank or defer interpretation when the user marks a major confound such as facilitator role, structured low-speaking role, or voluntary low-participation preference.

Why abstention is a feature. In a governance product, abstention increases trust because it shows the system knows where its boundaries are. Forced inference in weak cases makes the entire product easier to reject.

3.3 Detector taxonomy Kashi should adopt

To stop the current "no content" confusion, Kashi should explicitly classify detectors into tiers. Recommended taxonomy:

Tier	Definition	Examples for Kashi
Tier 1: Structural	Derivable from timing, speaker identity, overlap, adjacency, and participation counts alone.	Speaking-share inequality, overlap interruption, latency, dyadic concentration, chilling-delta.
Tier 2: Constrained transcript-semantic	Rule-based or fixed-model processing of transcript text without open-ended employer-side content interpretation.	Ignored-turn/topic-credit, unanswered-question logic, limited agreement-shift logic.
Tier 3: Forbidden / out of scope	Claims the system should refuse to make.	Emotion inference, intent inference, harassment classification, voice stress, future behavior prediction.

4. Exact project-useful changes to the current Kashi materials

4.1 Sentences to weaken or replace

Current phrasing	Problem	Recommended replacement
“The pattern is the harm.”	Too strong; equates observed pattern with moral/legal reality.	“The pattern may constitute evidence consistent with uneven conversational treatment over time.”
“Detecting harmful team dynamics earlier.”	Sounds like direct harm detection.	“Surfacing repeated interaction asymmetries earlier.”
“Never transcribe for analysis / none read meeting content.”	Conflicts with multiple detector descriptions.	Either remove those detectors from MVP or rewrite as “core detection is structural-first; limited constrained transcript-semantic detectors may support selected signals.”
“How we know we’re right.”	Overstates current validation stage.	“What the current pilot demonstrates.”
“All deterministic” where embedding or semantic inference is involved	Overstates epistemic purity.	“Rule-based, repeatable, and explainable; no live generative inference on the detection path.”

4.2 Replacement paragraph for the scientific-foundation section

> *Kashi does not detect harm, determine misconduct, or infer intent. It estimates whether repeated meeting interactions display context-conditioned asymmetry in floor access, interruption directionality, response burden, and participation change relative to appropriate baselines. These outputs are review-support signals rather than findings. Their interpretability depends on comparable exposure, input quality, detector scope, and unresolved confounds. When the evidence basis is weak, the system should abstain or present the signal as provisional.*

4.3 Add a new explicit subsection: “Measurement science and epistemic limits”

The current deck has the ingredients for this section but not the section itself. Add a compact subsection with seven points:

- Construct — Kashi estimates interaction asymmetry, not harm itself.
- Baselines — outputs depend on self-history, within-meeting, meeting-type, role, and dyad baselines.
- Comparability — scores are not assumed fully comparable across teams, cultures, languages, or meeting genres without evidence.
- Credibility — person-level interpretation requires repeated comparable exposure; meeting count alone is insufficient.
- Uncertainty — outputs carry evidence grade, reason codes, and abstention where needed.
- Input quality — transcript and speaker-ID limitations directly constrain downstream inference.
- Use boundary — outputs support human review and reflection; they do not determine misconduct, discipline, or legal status.

5. Pilot validation roadmap the project can actually execute

This section is written to be actionable for the team. It is designed to convert the research result into a staged work plan rather than an abstract critique.

Stage	Goal	What to collect	What success looks like	Why this matters
Stage 1 — Mechanism sanity	Verify that authored patterns fire as expected and healthy controls stay clean.	Synthetic scenarios, regression tests, detector traces.	Stable reruns, clean event traceability, obvious bug reduction.	Necessary but not sufficient.
Stage 2 — Confound testing	Stress-test introversion, facilitator role, L2, sparse participation, and rough-but-benign disagreement.	Synthetic + semi-synthetic adversarial cases.	Detector behavior remains conservative; abstention works.	Shows the system can fail gracefully.
Stage 3 — Input-quality audit	Measure behavior under transcript/diarization degradation and overlap-heavy sessions.	Real meeting samples with quality annotations; ASR confidence logs.	Low-quality inputs are visibly down-ranked or blocked.	Prevents pseudo-evidence from weak input.
Stage 4 — Human-review alignment	See whether independent reviewers judge surfaced patterns as worth review.	Blind reviewer study on event bundles, not just raw clips.	Moderate+ agreement on "review-worthiness," not necessarily on moral labels.	This is closer to the actual product use.
Stage 5 — Small pilot monitoring	Observe whether the tool is trusted, challenged, ignored, gamed, or useful in practice.	Usage logs, challenge rates, interview feedback, incident notes.	Useful without clear misuse; visible anti-gaming adjustments.	Pre-deployment eval alone will not answer this.
Stage 6 — Governance refinement	Convert pilot findings into retention, access, and reporting rules.	Policy decisions, audit results, product changes.	Sharper boundaries and a clearer deployment story.	The governance product is the product, not only the detector.

5.1 What success should mean during the pilot

The pilot should not define success as "the score moved." That is too easy to game and too weak scientifically.

Better pilot success criteria:

- The surfaced events are understandable and traceable to concrete turn evidence.
- Users can tell when a signal is weak, provisional, or confounded.
- Employees do not experience the product as broad managerial surveillance.
- Managers do not treat the outputs as quasi-disciplinary truth.
- Challenge/appeal pathways are used and actually improve the record.
- The system can gracefully say "not enough evidence" in ambiguous cases.
- Pilot logs reveal where signals are displaced or gamed, not only where they improve.

6. Immediate action list for the project team

- Decide within the next deck revision whether Kashi is structural-only MVP or honest hybrid MVP. Do not leave the current content contradiction unresolved.
- Replace all deck lines that imply direct harm detection with interaction-asymmetry language.
- Add a measurement-science / epistemic-limits subsection to the deck and website.
- Operationalize confidence as evidence grade + reason codes + abstention triggers rather than a hidden factor inside one composite score.
- Downgrade the evaluation claim from “how we know we’re right” to “what the current pilot demonstrates.”
- Build a confound/adversarial test pack: facilitator-heavy meeting, rough but normal design review, low-speaking SME meeting, multilingual meeting, overlap-heavy audio meeting, and sparse-data meeting.
- Add transcript-quality and diarization-quality logging to the pipeline before treating output as pilot-grade.
- Prepare one investor/judge slide titled “What Kashi measures / what Kashi refuses to claim.” This single slide will reduce avoidable argument loss.

If the team does only three things now… 1) fix the content contradiction, 2) narrow all rhetoric to asymmetry-under-uncertainty, and 3) add evidence grades plus abstention. Those three changes alone make the project materially harder to dismiss.

7. Final judgment

Kashi is closest to being credible when it is modest. Its real strength is not that it can “know” workplace harm from transcripts. Its real strength is that it can make repeated interaction asymmetries visible enough that an institution can no longer plausibly treat every meeting as an isolated fragment.

That is already a strong product thesis. It does not need overclaiming. In fact, overclaiming is what would weaken it. The research result therefore points to a disciplined strategic conclusion: Kashi should lead with bounded measurement, traceable signals, and explicit uncertainty. That makes the project more serious technically, more defensible legally, and more believable commercially.

Said plainly: Kashi should present itself not as a truth machine for workplace harm, but as governance infrastructure for making possible patterns of uneven conversational treatment visible early enough to review, contest, and act on appropriately.

Selected sources used for this memo

- Internal project source: “Kashi — Progress & Project Overview (2026-04-21).” Uploaded project PDF used as the primary object of critique and tailoring.
- Internal project source: “Transparency That Drives Institutional Accountability” concept note (meeting governance concept note). Used for alignment with the longer-form concept framing.
- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. Standards for Educational and Psychological Testing. 2014. Official PDF: testingstandards.net.
- National Institute of Standards and Technology. AI Risk Management Framework (AI RMF 1.0). NIST AI 100-1. 2023. Official PDF: nvlpubs.nist.gov.
- National Institute of Standards and Technology. Practices for Automated Benchmark Evaluations of Language Models and AI Agent Systems. NIST AI 800-2 ipd. January 2026. Official PDF: nvlpubs.nist.gov.
- National Institute of Standards and Technology. Expanding the AI Evaluation Toolbox with Statistical Models. NIST AI 800-3. 2026. Official PDF: nvlpubs.nist.gov.
- National Institute of Standards and Technology. Challenges to the Monitoring of Deployed AI Systems. NIST AI 800-4. March 2026. Official PDF: nvlpubs.nist.gov.
- Koenecke, A. et al. “Racial Disparities in Automated Speech Recognition.” Proceedings of the National Academy of Sciences, 2020. Official article: pnas.org/doi/10.1073/pnas.1915768117.
- Mujtaba, D. et al. “Lost in Transcription: Identifying and Quantifying the Accuracy Biases of ASR Systems Against Disfluent Speech.” NAACL / ACL Anthology, 2024. Official entry: aclanthology.org/2024.naacl-long.269/.
- Köhler, T. & Gölz, M. “Meetings Across Cultures: Cultural Differences in Meeting Expectations and Processes.” In The Cambridge Handbook of Meeting Science, 2015. Cambridge University Press.
- Kreamer, L. M. et al. “Virtual Voices: Exploring Individual Differences in Written and Verbal Participation in Virtual Meetings.” Organizational Behavior and Human Decision Processes, 2024. Microsoft Research PDF available online.
- Shavelson, R. J., Webb, N. M., & Rowley, G. L. “Generalizability Theory.” American Psychologist, 1989/1990 reprint PDF widely circulated. Used here for the dependability-of-measurement logic.
- Center for Naval Analyses. Goodhart’s Law: Recognizing and Mitigating Manipulation of Measures in Analysis. 2022. Official PDF: cna.org.