Kashi Scientific Attack-Surface Memo

Project-use memo tailored to the current Kashi progress deck and concept note

Prepared for: Kashi project team • Primary use: deck revision, pilot design, hostile-review preparation, and validation planning • Date: 2026-04-21

Bottom line. Kashi is strongest when it treats the current detector set as a bounded, uncertainty-aware review system rather than a truth machine. The biggest scientific attack vectors are not whether asymmetric interaction exists as a phenomenon, but whether the current detectors are equally justified, whether sparse exposure is being over-read, whether transcript / diarization noise is contaminating outputs, and whether the deck is pretending that 'confidence' is already defined when it is not. The cleanest project move is to add a deliberate 'where this can break' section and make that discipline part of the product story itself.

Executive conclusion

- The current six-detector set is directionally sensible because it concentrates on comparatively observable properties of meetings — interruption, floor access, response burden, and repeated directionality — rather than on tone, intent, or psychologized claims. But the six detectors are not equally strong. Intrusive interruption, floor-share inequality, and repeated directional concentration are materially more defensible than topic-credit recovery or agreement-asymmetry.
- The current k≥5 meetings / ≥30-day rule in the progress deck is best treated as a cold-start floor, not as a signal-stability proof. In analogous observational work using Generalizability Theory, stable estimates can require roughly five to more than fifteen occasions depending on the variable; meeting count alone is not the right criterion.[5]
- Topic-credit detection is the scientifically weakest named detector in the current deck. The underlying phenomenon is real, but automation requires transcript semantics, paraphrase matching, and credit assignment logic that are much less mature and much easier to attack than pure structural timing signals. Pairwise semantic tasks are known to generalize poorly under realistic class imbalance, and meeting-language semantics differ substantially from clean article-style text.[10]
- Deterministic pipelines still drift. Even without live generative inference, Kashi can drift because upstream transcription/diarization changes, meeting types change, org structure changes, norms change, and per-speaker baselines go stale. Calibration freshness therefore has to be treated as an explicit monitoring problem, not as something 'deterministic' magically solves.[3][4][7]
- ASR and diarization failure are a first-order scientific risk. In meeting benchmarks, speaker-attributed transcription remains substantially worse than close-talk transcription and degrades sharply under overlap; overlap-heavy recordings can show error rates dramatically above low-overlap conditions.[8][9][11][12] Kashi must therefore gate outputs on input quality.
- In a deterministic pipeline, 'confidence' should not mean model probability. For Kashi it should mean evidence adequacy: input quality, comparable exposure, unresolved confounds, detector coherence, and calibration freshness, expressed as an evidence grade plus reason codes and abstention triggers rather than a hidden multiplier inside one composite score.[2][3][4]

1. What this memo is trying to do

This memo is not another generic 'measurement science' explainer. It is a critic-facing attack memo built for project use. The target question is narrower and harsher: if a skeptical judge, reviewer, CISO,

counsel, or technically literate HR lead wanted to stress-test the current Kashi progress deck, where would they push, and which pushes would actually land?

The current progress deck already does something unusually right: it narrows the claim. It explicitly rejects harassment classification, affect inference, automatic action, HR-decision use, and company-wide health scores; it uses review-worthy events instead of verdict language; and it calibrates to a speaker's own 90-day baseline instead of team average.[I1] That is already much better than the average workplace-AI pitch.

The real problem now is not that the concept lacks scientific caution. The problem is that the caution is uneven. Some parts of the deck are careful; other lines still drift toward stronger rhetoric ('the pattern is the harm', 'detecting harmful team dynamics earlier', 'all deterministic', 'none read meeting content'). A hostile reviewer will always pull on the strongest-sounding line, not on the most careful one. The job of this memo is therefore to tighten the weak joints on purpose.

Hostile reviewer questions: short answers Kashi should be ready to give

Hostile question	Best disciplined answer	Why this answer survives better
Why these six detectors and not others?	Because they mostly target lower-inference, meeting-observable structure: floor access, interruption, response burden, and repeated directionality. But they are not equally strong; topic-credit and agreement-asymmetry should be framed as higher-fragility or experimental tiers, not as co-equal truths.	It defends the design logic without pretending all six have the same evidentiary status.
How many meetings before a signal is stable?	There is no single magic number. Stability depends on comparable exposure, interaction opportunity, and input quality. The current k≥5 / 30-day rule is a reporting floor, not a validity proof; analogous observational work shows some variables need roughly 5 to >15 occasions for stable estimation.[5]	It avoids fake precision while still giving a research-backed structure.
How robust is topic-credit detection?	Conceptually real, operationally fragile. It requires transcript semantics, paraphrase matching, and credit attribution logic that are materially weaker than pure timing-based signals. In MVP, it should be experimental or user-side assistive, not a core institutional score.[10]	It shows maturity instead of pretending a hard NLP problem is already solved.
What does confidence mean if the pipeline is deterministic?	Not model probability; evidence adequacy. Confidence should aggregate input quality, comparable exposure, unresolved confounds, detector coherence, and calibration freshness into evidence grades and abstention triggers.[2][3]	It turns a vague word into a concrete measurement layer.

2. Why these six detectors — and why they are not equally defensible

The best defense of the current detector set is not 'these are the six true indicators of workplace harm.' That would be brittle and wrong. The better defense is that the deck selects a first-pass family of signals that are comparatively accessible from timestamped meeting traces and speaker attribution

without crossing into prosody, emotion, visual surveillance, or open-ended moral classification. That is already a valid design principle.

In other words, the current set is defensible as a low-inference starting family, not as an exhaustive ontology. A hostile reviewer will immediately ask why these six and not others. The right answer is: because Kashi is deliberately prioritizing what is more observable, repeatable, and governable from meeting records alone — and refusing many things that would require richer modalities, higher inference, or legally hotter claims.

That same answer also implies a necessary concession: the six are not equally strong. Some are close to direct meeting-structure measures. Others depend on transcript semantics, turn segmentation assumptions, or paraphrase-like comparisons and should therefore be explicitly tiered.

Detector-by-detector attack assessment

Detector	Why it is directionally sensible	Main attack surface	Recommended status	Risk
Intrusive interruption	Strong fit to turn-level structure; comparatively low semantic dependence; well aligned with dominance / interruption literature already cited in the deck.	Turn segmentation and overlap detection can still fail under noisy audio, but the construct itself is comparatively clean.	Keep as Tier 1 structural.	Low–medium
Chilling delta	Captures post-event participation change relative to own baseline; stronger than one-off silence counts because it is directional and personal-baseline aware.	Sensitive to agenda changes, timing windows, sparse speech, and mistaken trigger events.	Keep as Tier 1 structural, but show caveats.	Medium
Floor-time Gini	Useful descriptive inequality measure for floor access; good dashboard metric when treated descriptively, not morally.	Easy to over-interpret; domination can be role-driven or meeting-type driven rather than suppressive.	Keep as descriptive Tier 1 metric only.	Medium
Unanswered-question rate	Plausible response-burden signal if question detection and substantive-response logic work reasonably.	Requires transcript interpretation; 'substantive response' is already more semantic than the deck admits.	Move to Tier 2 constrained transcript-semantic.	Medium–high
Topic-credit ignored-turns	Real social phenomenon and strategically important for Kashi's thesis.	Hardest named detector to defend: requires paraphrase matching, idea equivalence, and credit attribution across transcript noise and class imbalance.[10]	Experimental Tier 2 only; not core institutional score in MVP.	High
Agreement	Interesting as a	Weakest evidentiary grounding	Prototype / research-	High

Detector	Why it is directionally sensible	Main attack surface	Recommended status	Risk
asymmetry	pressure / convergence hypothesis.	operationally; position shift is hard to infer from transcript snippets and may require richer discourse or role modeling.	only until validated.	

Why not more detectors?

The clean answer is that several additional detector families would either violate Kashi's current legal/governance boundaries or require richer modalities and more fragile inference than the present deck can honestly support. Prosody, affect, and emotion-style claims are explicitly out of scope and, in some workplace contexts, legally red-lined. Role and addressee inference are also much harder than they look; recent multimodal work shows that high-quality speaker/addressee/role attribution often needs full audio-visual context, not just text, and degrades as group complexity rises.[13]

This means Kashi should defend omission as discipline, not as absence of imagination. The platform is stronger when it says: we chose signals that are more compatible with structural meeting records and bounded governance; we did not choose the signals that would require body-language analytics, psychologized emotion inference, or broad semantic interpretation to work.

3. Signal stability: minimum meeting count is the wrong primitive

The current Kashi deck already says that a single meeting is noise and that cross-meeting windows matter. Good. But the current k≥5 meetings / ≥30-day line is easy to misread as 'after five meetings we know.' That is not scientifically safe. The more defensible claim is narrower: five comparable meetings may be enough to begin surfacing provisional patterns, but not enough to guarantee stable person-level inference.

Generalizability Theory is useful here because it treats repeated behavioral observation as a dependability problem with multiple sources of variance — occasion, context, role, speaker mix, and other facets — rather than as one simple sample-size issue.[5] In one observational study that explicitly used G-Theory to plan sampling, estimates for different language variables required roughly five to more than fifteen occasions to reach a reliability threshold of 0.80.[5] The lesson is not that Kashi should hard-code '15 meetings.' The lesson is that stability is variable-specific.

For Kashi, the correct unit is comparable exposure, not raw meeting count. Three high-interaction sprint reviews may be more informative than ten passive all-hands meetings. A person with repeated dyadic exchanges yields more evidence than a low-speaking attendee. A one-off customer call should probably contribute little or nothing to person-level inference at all.

What Kashi should say instead of a magic meeting count

Weak answer to avoid	Stronger answer Kashi should give
"After 5 meetings we know."	"Credibility depends on repeated comparable exposure, interaction opportunity, and input quality. The current k≥5 / 30-day rule is a reporting floor; stable interpretation becomes stronger only when the evidence basis deepens."

4. Topic-credit detection: real phenomenon, fragile automation

This is probably the cleanest scientific attack point in the current deck. The underlying social phenomenon is not the issue: people do get ignored, restated, and de-credited. The issue is whether current meeting-transcript automation can robustly identify that phenomenon at production quality without overclaiming.

The current deck describes topic-credit ignored-turns as 'A proposes → ignored → B restates similar content → B is credited', detected via a turn-similarity graph.[I1] The problem is that every element after 'A proposes' is fragile. 'Similar content' is already a paraphrase / semantic-similarity task. 'Ignored' is not a purely structural fact; it is partly discourse interpretation. 'Credited' may be explicit, implicit, delayed, or role-mediated. And all of this sits on top of transcript and diarization noise.

This concern is not hand-wavy. Pairwise semantic tasks such as paraphrase identification are known to generalize poorly when deployed on realistically imbalanced data rather than tidy benchmark distributions.[10] Meeting-transcript language is also known to differ substantially from article-style or clean written text, which weakens naive transfer from standard semantic benchmarks to long, messy meetings.[10] That does not mean topic-credit is impossible. It means the detector should be framed as high-fragility unless Kashi invests in a serious validation program with annotated real meeting data.

Project-use judgment. Topic-credit should not be sold as a core hard detector in the same breath as interruption or floor-share. In the current product state it is better framed as: (a) an experimental Tier 2 detector, (b) a victim-side assistive cue, or (c) a human-review accelerant that suggests where a reviewer should look — not as a stable institutional score.

5. Calibration drift under org change: determinism does not save you

A deterministic pipeline can still drift. 'Deterministic' only means the same input returns the same output. It does not mean the measurement remains equally calibrated as the environment changes. Kashi is vulnerable to at least five drift channels: upstream platform ASR/diarization changes; meeting-format changes (incident reviews, retros, layoffs, hiring spikes, new manager cadence); org-structure changes that alter role expectations; language-environment changes; and baseline staleness when a person's recent history no longer represents the current regime.

NIST’s AI RMF and its newer evaluation and monitoring documents push exactly this point in broader form: claims should be qualified, uncertainty should be reported, and post-deployment monitoring is necessary because conditions change and pre-deployment tests are not enough.[2][3][4] Kashi should therefore stop letting 'deterministic' carry an implicit meaning of 'stable.'

The minimum project response is to treat calibration freshness as a measurable state. A baseline computed before a reorg, before a new manager arrives, or before the team shifts languages may be mathematically neat and scientifically stale.

What Kashi should monitor for drift

Drift source	Why it changes the meaning of the score	Minimum monitoring response
Upstream transcription / diarization changes	Same speaking behavior may produce different text or speaker-attribution quality after platform model updates.	Version logging, shadow comparison set, alert on major quality shifts.
Meeting-type shift	Standups, retros, all-hands, and incident calls are not behaviorally interchangeable.	Meeting-type tagging and type-specific baselines.

Org / role change	New manager, reorg, or facilitation change alters who is expected to speak or direct discussion.	Baseline reset / freshness checks after major org changes.
Language mix change	L2 burden, multilingual switching, and dialect shifts can change latency and floor-share patterns.	Language-context flag and stronger abstention under mixed-language conditions.
Behavioral adaptation / gaming	People may optimize visible metrics while preserving underlying dominance in other channels.	Post-deployment monitoring, dispute review, and anti-gaming audits.

6. ASR / diarization poisoning: where outputs can be contaminated before detector logic starts

This is the most practical scientific failure mode because Kashi starts from transcript-linked meeting traces and speaker attribution. If that input layer is wrong, the downstream detector can be internally consistent and still scientifically misleading.

Meeting transcription remains difficult in precisely the conditions Kashi cares about: spontaneous multi-party speech, overlap, turn changes, and real-world audio. In the NOTSOFAR-1 meeting transcription challenge, baseline speaker-attributed tcpWER on all sessions was 32.4% for multichannel and 46.8% for single-channel systems, worsening to 38.0% / 54.9% on sessions labeled DebateOverlaps.[8] In another analysis of conversational recordings, MIMO-WER ranged from 5.99% on mostly single-speaker recordings to 42.31% on recordings with significant overlap.[9]

The poisoning risk is not only average error. It is structured error. Work on lexical speaker error correction notes that diarization and reconciliation are especially error-prone around speaker turns and overlap, where words from one speaker may be misattributed to another.[11] Work on disfluent speech shows statistically significant ASR accuracy bias against disfluent speech, with degraded syntactic and semantic accuracy in the resulting transcript.[12] And widely cited fairness audits of commercial ASR showed substantial disparities across speaker groups, including average WER around 0.35 for Black speakers versus 0.19 for white speakers across five systems.[6]

The implication for Kashi is brutal but simple: some proportion of the signal may be measuring who the speech system fails on, not who the team is treating unevenly. A product that does not surface this limitation will look unserious under scientific scrutiny.

Minimum gating rules Kashi should add before calling the output pilot-grade

- Transcript-quality logging at the segment and meeting level.
- Diarization-quality / speaker-attribution confidence logging before any directional detector is trusted.
- Overlap-burden flag that can down-rank interruption / chilling interpretations in severe crosstalk.
- Language / L2 / multilingual caution surface rather than silent uniform scoring.
- Pilot subgroup audit that checks whether error rates cluster by speech style, disfluency, accent, or meeting format.

7. What 'confidence' should mean in Kashi’s deterministic pipeline

Right now 'confidence' appears inside the composite score formula in the progress deck, but it is not yet operationalized.[I1] That makes it scientifically vulnerable because it looks like a hidden stabilizer rather than a principled part of the measurement model.

In Kashi's context, confidence should not mean model probability. The better concept is evidence adequacy. This is much closer to measurement science and much more defensible in front of a hostile reviewer: how good is the input basis, how much comparable exposure exists, how many confounds remain unresolved, how coherent are the detectors, and how fresh is the calibration regime?[2][3][4]

The best product move is therefore to turn confidence into a user-visible evidence grade rather than keeping it as a silent scalar buried inside one formula.

Evidence component	What it should include
Input quality	Transcript confidence, diarization quality, overlap burden, language-context warning.
Exposure adequacy	Number of comparable meetings, number of relevant exchanges, dyadic recurrence, not just raw calendar count.
Confound state	Facilitator role, L2 status, sparse-speaking role, voluntary low-participation preference, unusual meeting format.
Detector coherence	Whether multiple detector families indicate the same direction rather than one fragile detector firing alone.
Calibration freshness	Whether the baseline predates a reorg, new manager, or meeting-regime shift.

Recommended output ladder. Insufficient evidence → Weak pattern → Emerging pattern → Stable pattern. Each state should carry reason codes such as low transcript quality, sparse comparable exposure, unresolved confounds, or stale baseline. This is materially stronger than a single naked score.

8. Paste-ready section for the deck: "Where Kashi can break"

If the team wants to pre-empt attack instead of merely surviving it in Q&A, the progress materials should add a direct section named something like 'Where Kashi can break' or 'Scientific limits and failure modes.' This should not read like an apology. It should read like disciplined system design.

Suggested paragraph for direct reuse

Kashi does not assume that every detector is equally stable in every meeting context. Some signals, such as interruption concentration or floor-access inequality, are closer to direct structural observation; others, such as topic-credit recovery or agreement asymmetry, are more fragile because they depend on transcript interpretation, semantic similarity, or higher discourse assumptions. The system is also sensitive to input quality: poor transcription, diarization error, overlap-heavy audio, multilingual conditions, sparse comparable exposure, and stale baselines can all weaken interpretability. For this reason, Kashi should treat outputs as review-support signals under uncertainty, expose evidence grades and reason codes, and abstain when the basis is weak rather than forcing a confident-looking score.

Exact wording changes the current deck should make

Current phrasing or move	Recommended correction
“The pattern is the harm.”	Replace with: “The pattern may provide evidence consistent with uneven conversational treatment over time.”
“Detecting harmful team dynamics earlier.”	Replace with: “Surfacing repeated interaction asymmetries earlier.”
“All deterministic / none read meeting content.”	Either (A) remove transcript-semantic detectors from MVP, or (B) rewrite as: “core detection is structural-first; selected constrained transcript-semantic detectors may support narrower signals.”
“Confidence” as a hidden score factor only.	Replace with visible evidence grades, reason codes, and abstention triggers.
“How we know we’re right.”	Downgrade to: “What the current pilot demonstrates.”

9. P0 / P1 actions for the project

Priority	Action	Why this matters
P0	Tier the detectors explicitly into structural vs constrained transcript-semantic vs forbidden.	Stops the current 'metadata only' contradiction from silently eroding credibility.
P0	Operationalize confidence as evidence grade + reason codes + abstention triggers.	Turns a vague multiplier into a real measurement layer.
P0	Add transcript-quality and diarization-quality logging before treating output as pilot-grade.	Without input-quality gating, downstream asymmetry signals are easier to attack.
P0	Rewrite the scientific foundation and hero rhetoric to asymmetry-under-uncertainty language.	Prevents hostile reviewers from quoting stronger wording back at the team.
P1	Build a confound / adversarial validation pack: facilitator-heavy meeting, rough-but-normal design review, sparse-speaking SME meeting, multilingual meeting, overlap-heavy meeting, reorg month.	The current seed scenarios prove mechanism, not robustness.
P1	Define baseline-reset and calibration-freshness rules after manager change, reorg, or meeting-type shift.	Deterministic baselines still drift when the institution changes.
P1	Demote topic-credit and agreement-asymmetry in public scoring until real-world validation exists.	These are the easiest named detectors for critics to attack first.

10. Final judgment

Kashi is already stronger than most adjacent workplace-AI concepts because it is willing to narrow the claim, refuse obvious poison features, and frame outputs as review-worthy rather than dispositive. That is real strength, not defensive weakness.

The current scientific attack surface is also real. The most vulnerable areas are detector equality, topic-credit robustness, sparse-exposure over-reading, input-layer contamination, and undefined confidence. None of those kill the project. But if they remain implicit, critics will get to name them first.

The project therefore does not need a bigger claim. It needs a sharper boundary. The serious version of Kashi is not: 'we detect harmful power dynamics.' The serious version is: 'we surface repeated interaction asymmetry under uncertainty, expose when the basis is weak, and refuse to pretend the system knows more than meeting records can honestly support.' That is harder to attack technically, cleaner to defend in a deck, and much more believable in a real pilot.

Selected sources

[I1] Kashi — Progress & Project Overview (2026-04-21). Internal project PDF used as the primary object of critique and tailoring.

[I2] Kashi Measurement-Science Research Memo (2026-04-21). Internal project memo used as the immediate conceptual baseline for this follow-on attack-surface brief.

[1] American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. Standards for Educational and Psychological Testing. 2014.

[2] NIST. Artificial Intelligence Risk Management Framework (AI RMF 1.0). NIST AI 100-1. 2023.

[3] NIST. Practices for Automated Benchmark Evaluations of Language Models and AI Agent Systems. NIST AI 800-2 ipd. January 2026.

[4] NIST. Challenges to the Monitoring of Deployed AI Systems. NIST AI 800-4. March 2026.

[5] Ford, A. L. B., et al. 'The Use of Generalizability Theory to Inform Sampling of Educator Language Used With Preschoolers With Autism Spectrum Disorder.' Journal of Early Intervention, 2021. Key project-use finding: some observed variables required roughly five to more than fifteen occasions for stable estimates.

[6] Koenecke, A., et al. 'Racial Disparities in Automated Speech Recognition.' Proceedings of the National Academy of Sciences, 2020.

[7] NIST. Expanding the AI Evaluation Toolbox with Statistical Models. NIST AI 800-3. February 2026.

[8] Vinnikov, A., et al. 'NOTSOFAR-1 Challenge: New Datasets, Baseline, and Tasks for Distant Meeting Transcription.' Interspeech 2024. Baseline speaker-attributed tcpWER on all sessions: 32.4% (MC) / 46.8% (SC), worsening in overlap-heavy conditions.

[9] Maciejewski, M., et al. 'Evaluating the Santa Barbara Corpus: Challenges of the Last-Fifty-Years Data for Modern ASR.' Interspeech 2024. Reported MIMO-WER ranging from 5.99% on mostly single-speaker recordings to 42.31% on recordings with significant overlapped speech.

[10] Mussmann, S., et al. 'On the Importance of Adaptive Data Collection for Extremely Imbalanced Pairwise Tasks.' Findings of EMNLP 2020. Key project-use point: state-of-the-art pairwise models generalized poorly on realistically imbalanced paraphrase-like tasks.

[11] Paturi, R., et al. 'AG-LSEC: Audio Grounded Lexical Speaker Error Correction.' 2024. Key point: diarization and reconciliation are especially error-prone around speaker turns and overlapping speech.

[12] Mujtaba, D., et al. 'Lost in Transcription: Identifying and Quantifying the Accuracy Biases of Automatic Speech Recognition Systems Against Disfluent Speech.' NAACL 2024.

[13] Rayan, J. A., et al. 'Multimodal Conversation Structure Understanding.' 2025/2026 preprint. Project-use point: high-quality role/addressee attribution benefits materially from full audio-visual context, not just text.