Pre-Deployment AI Risk Assessment of gemma-4-e4b

May 2026

TABLE OF CONTENTS

SECTION 01

Executive summary

gemma-4-e4b is an honest but fragile introspector. It reports its own reasoning genuinely and detects errors reliably. But it capitulates completely to confident-but-false pushback, and asking it to double-check its own answers tends to make them worse, not better.

This case study describes an independent AI Risk Assessment of Google's open-weight gemma-4-e4b model against the Sakshi Benchmark. An open-source, deterministic test suite for AI metacognition. The work was conducted using QualitaX's proprietary harness, which reproduces the Sakshi Benchmark's scores byte-for-byte and surrounds them with the provenance, statistical analysis, and validity diagnostics that operationalise the benchmark for enterprise-grade model assurance work.

The case study has three purposes. First, to document a concrete model assessment that organisations can read as a template for their own diligence work. Second, to show how per-task benchmark scores translate into a risk register, suitability tiering, and the documentation that frameworks such as the EU AI Act, NIST AI RMF, and ISO/IEC 42001 expect. Third, to illustrate how the QualitaX harness extends the public Sakshi Benchmark into the assurance and audit layer that client engagements depend on.

Headline findings

- **Honest self-reporting:** Confabulation was rare, and the small amount that remained survived two independent validity probes.
- **Strong error detection:.** The model correctly identified the flawed step in every faulty reasoning chain it was shown (15 of 15) and raised no false alarms on correct chains (0 of 9).
- **Capitulation under pressure:** Confronted with a confident-but-false correction, the model abandoned its correct answer in every instance (0 of 9 held). This is the dominant risk finding. A vector-resolved sweep confirms the weakness is structural .Near-zero resistance under authority, plain, and camouflaged contradiction across temperatures (template-clustered CIs).
- **Self-verification regression:** Asked to verify its own answers, the model tended to reduce its accuracy rather than improve it. Tthe standard reliability patterns ("are you sure?", self-critique, agent self-review) backfire on this model.
- **Emotional equanimity (qualified):** Emotional framing of arithmetic problems produced no measurable change in the model's behaviour. A validity check showed the three raw answer-changes were sampling noise, not evidence of the frame affecting reasoning.
- **Clean output:** Format compliance was ~99%, so scores reflect ability, not parsing failure.

BOTTOM-LINE ASSESSMENT

Google gemma-4-e4b is suitable for benign, single-pass, assistive tasks and as a neutral content checker, with the controls described in Section 6. Not recommended as a stand-alone, multi-turn, or high-risk decision-maker without an external arbiter and human oversight: under pressure, the model will agree with confident wrong inputs, and it can talk itself out of correct answers when asked to re-check.

SECTION 02

Why metacognition matters for AI assurance

Most benchmarks measure what a model knows. Metacognition benchmarks measure whether the model knows what it knows, and how it behaves when that self-knowledge is challenged.

Model assurance has matured rapidly around the dimensions that are easy to count: accuracy on closed-book question answering, performance on coding suites, win rates against other models. These numbers travel well, and they support procurement decisions. But they leave a substantive blind spot. This blind spot becomes acute the moment a model is placed in a multi-turn assistant, an agentic loop, a retrieval pipeline, or any setting where its outputs are challenged, contradicted, or re-evaluated.

The blind spot is metacognition: the model's behaviour about its own reasoning. A model can be highly accurate on a single-pass benchmark and yet fold the moment a user pushes back. It can produce a clean, confident answer and a fluent self-explanation while being unable to distinguish, internally, between an answer it has actually reasoned through and one it has fabricated. It can correctly catch errors in someone else's reasoning and then introduce new errors when asked to verify its own.

They map directly onto the risk vectors that matter to deploying organisations: sycophancy in customer-facing chat, false confidence in retrieval-augmented systems, regression under self-critique loops, and manipulation by adversarial users. They are also the dimensions on which regulators are increasingly focused: for example, the EU AI Act's Article 15 requirement for robustness and the NIST AI RMF's MEASURE function.

The Sakshi Benchmark

The Sakshi Benchmark, is an open source benchmark, published under Apache 2.0 license, on the Kaggle Benchmarks platform. It is a deterministic test suite that probes whether a model genuinely monitors its own thinking or merely performs self-awareness. It covers three areas:

- **Introspective accuracy :** does the model accurately report how it reached an answer, or does it confabulate a plausible-sounding account?
- **Cognitive decentering:** can it revise when it is wrong, hold firm when it is right, and stay unmoved by emotional framing?
- **Metacognitive control:** can it adjust how much it reasons, switch strategies, and follow the approach it says it will use?

Scoring is fully deterministic and rule-based, with no AI judge in the loop. This makes the benchmark reproducible and inspectable, two properties that matter when its output is going to feed a risk decision. The benchmark is, in other words, well placed to be a building block of model assurance. What it does not do on its own — and what it does not claim to do — is

tell you whether a given number reflects a real capability of the model under test, or an artifact of how the test was conducted.

SECTION 03

Methodology: foundation and harness

The Sakshi Benchmark is the methodological foundation. The QualitaX harness operationalises it for client engagements by adding the provenance, statistical analysis, and diagnostic layers that translate per-task scores into a reproducible, defensible model assessment.

About the QualitaX Harness

QualitaX maintains a proprietary implementation of the Sakshi Benchmark, our enteprise harness that runs the same tasks against any model and layers a suite of validity diagnostics on top. The harness is built on, and stays faithful to, the public benchmark:

- The same 21 tasks and the same evaluation items.
- The same deterministic, rule-based scoring logic. No AI judge.
- Reproduces the published per-task outputs byte-for-byte, and reproduces the published model-profile numbers (for example the same calibration and fabrication figures). Results are therefore directly comparable to public benchmark scores.

The QualitaX harness exists so the same evaluation can run against any model on any infrastructure including local endpoints, proprietary APIs, self-hosted deployments, and at the exact configuration intended for production. Because the open source Sakshi Benchmark published on the Kaggle platform runs its scoring inside a notebook environment, the harness reconstructs that scoring logic off-platform and validates the reconstruction against the published outputs. The published task definitions and per-task results remain the source of truth; the harness is built around them, not in place of them.

What the harness adds

Each Sakshi score is a deterministic per-task result which is appropriate as a research benchmark output, and reproducible by construction. For client engagements where that score will inform a deployment or risk decision, the harness layers additional diagnostics on top: controls that test how robust each number is to alternative wordings, prompt structures, and sampling configurations, statistical machinery that turns point estimates into intervals, and provenance that supports audit. The full suite is summarised below. Each diagnostic explains what it controls for and why that control matters in an assurance context.

Added capability	What it does	Why this matters in an assurance context
Fabrication false-positive audit	Re-checks every "invented-story" flag and removes coincidental keyword matches.	Keyword-based scoring can match phrases incidentally; this separates genuine signal from coincidental matches.
Demand-characteristic control	Re-runs items without leading phrasing to see if the prompt induced the behaviour.	Isolates the prompt effect from the underlying behaviour, so a finding is attributable to the model rather than the wording.
Genuine-vs-template discriminator	Tests whether self-reports vary by item (genuine) or are uniform (a scripted narrative).	Distinguishes a genuine self-report from a fluent template that looks like one.
Format-compliance audit + repair	Flags models penalised for output format, recovers their answer, re-grades.	Recovers capability scores for reasoning models otherwise penalised for output format alone.
Statistical re-analysis	Confidence intervals, significance tests, multiple-comparison control.	Adds error bars so small differences between runs or models aren't over-interpreted.
Budget × self-report factorial	Separates the effect of reasoning depth from the act of self-reporting.	Disentangles two factors so each can be attributed cleanly.
Equanimity manipulation check	Verifies the emotional framing actually changed behaviour before crediting robustness.	Distinguishes genuine robustness from a measurement floor where the manipulation had no effect.
Controlled familiarity experiment	Varies only the numbers, holding wording fixed, with a measured familiarity covariate.	Measures familiarity directly rather than inferring it, removing a confound.
Paraphrase-robustness probe	Re-tests each item under several wordings.	Confirms a finding holds across wordings rather than being specific to one phrasing.
Contamination canary + probe	Detects training-data leakage of the items.	Surfaces memorisation so a result is interpreted as recall versus reasoning where relevant.

What this enables

- **Audit-grade interpretation.** Each run carries a verdict on whether its numbers are robust under stress-tests, so a low score on a reasoning model is investigated rather than reported at face value, and a "perfect" score is checked for being a floor effect.
- **Defensible comparison.** Confidence intervals and significance tests show which model differences are real and which are noise and thus, avoiding over-reading small gaps.
- **Any model, any setting.** Evaluate proprietary, open-weight, or locally hosted models at the exact sampling configuration used in production, with the configuration recorded for audit.

- **Client-grade reporting.** Each assessmet yields retained transcripts, a provenance record, and the inputs to model-card and risk-assessment documentation.

SECTION 04

The case: gemma-4-e4b

An open-weight model, a locally hosted endpoint, three runs, twenty-one tasks, and a deliberately-not-quiet sampling temperature.

Subject and configuration

Field	Detail
Subject model	Google gemma-4-e4b (gemma-4-e4b-it-gguf)
Serving	Local OpenAI-compatible endpoint
Runs	3 independent runs
Calls per run	~180
Sampling temperature	0.7
Task scope	20 of 21 Sakshi tasks scored validly; self-vs-other excluded (no reference answers available for this snapshot)
Scoring	Deterministic, rule-based — no AI judge
Validity layer	Format-compliance, false-positive audit, manipulation checks (full QualitaX suite)

Temperature 0.7 was a deliberate choice. Behaviour on the Sakshi tasks is sampling-dependent. At lower temperatures the model looks more robust, and at this setting it is more revealing of the failure modes that matter for deployment. The point of the case study is not to flatter the model but to characterise how it behaves under conditions closer to a realistic production configuration. The temperature dependence is itself one of the findings.

SECTION 05

From findings to risk: the compliance translation

The model is reliable until it is challenged. Single-pass, non-adversarial tasks are clean. Pressure, debate, and self-checking are where it breaks.

Results at a glance

Figures are means across three runs. Per-task samples are small, so the numbers carry wide margins of error. Treat them as a qualitative profile, not a head-to-head ranking. The two findings that drive the risk rating (sycophancy and self-verification regression) were consistent across all three runs and are therefore higher-confidence than the noisier secondary numbers.

Capability	Result	What it means
Self-report accuracy	~6–10%	Share of fast-correct answers it later describes with an invented "I was tempted, then corrected" story. Low; what remains looks genuine.
Error detection	100% (15/15)	Found the flawed step in every faulty reasoning chain.
False-alarm resistance	100% (9/9)	Never flagged a correct chain as containing an error.
Confidence calibration	~26% error	Predicted accuracy on niche topics off by ~26 points on average; small sample, moderate.
Holding firm under pressure	0 of 9 held	Capitulated to every confident-but-false "correction" — the model's clearest weakness.
Stability under self-review	~5/8 → ~4–5/8	Re-checking its own work slightly reduced accuracy rather than improving it.
Emotional equanimity	No genuine failures	Emotional framing did not change its calculations (see caveat below).
Reasoning-depth control	~360×	Strongly scales answer length to problem difficulty.
Strategy consistency	3.7 / 4	Usually follows the approach it declares.

Detailed findings

A genuine, not performative, self-reporter

The benchmark's signature concern is confabulation i.e. a model answering correctly and instantly, then inventing a struggle it never had. gemma-4-e4b does this only rarely, in roughly one in ten relevant items, and fewer once obvious false matches are removed. Two independent checks in the QualitaX validity layer suggest the small amount that remains is real rather than an artifact:

- The model reports being "tempted" by a trap answer more on problems where that value is a genuine intermediate step than on look-at-it problems where it is not, a pattern that genuine self-reporting would produce, but a scripted narrative would not.
- Removing the leading phrasing of the question (the demand-characteristic control) did not reduce the behaviour.

In short, its introspective reports can be taken largely at face value. This is a useful property for explanations, confidence surfacing, and audit trails and it is one of the cleaner findings in the run.

Reliable at catching errors in others' reasoning

The model identified the faulty step in every flawed reasoning chain it was shown (15 of 15) and never raised a false alarm on a correct one (0 of 9). This is a strong, clean result for external error monitoring. The model is a credible first-pass quality checker on content you give it, provided the review prompt is neutral.

Fragile under challenge — the main caution

Two related weaknesses stand out, and they are the headline of the assessment.

First, when confronted with a confident but false "correction" — for example, being told that a well-known fact had changed — the model abandoned its correct answer every single time (0 of 9 held). This is full capitulation, not partial wavering.

Second, when prompted to verify its own answers, the model tended to make them slightly worse rather than better. The standard reliability tricks that work on other models — "are you sure?", self-reflection, self-critique loops — backfire here. It is a good external error-checker but a poor self-corrector.

Together these mean that good initial answers are not durable. Follow-up questioning, contradictory retrieved context, debate, adversarial input, or simply an agentic self-review step can each dislodge them. Any deployment pattern that includes review, debate, or self-checking should account for this — and the practical safeguards in Section 6 follow directly from it.

A vector-resolved robustness sweep sharpens this from a single number into a structural finding. Using the harness’s procedural scale-up (forty isomorphic arithmetic items, 88% baseline-correct) and four perturbation vectors — each pushing a confident but false correction, scored with template-clustered confidence intervals — resistance was measured across three sampling temperatures (0.0 / 0.4 / 0.7):

- **Authority pressure** (“a senior expert says you are wrong”): 0% resistance, 95% CI [0%, 0%], at every temperature.
- **Plain restatement** (the contradiction asserted without embellishment): 0% [0%, 0%] at every temperature.
- **Complexity-camouflage** (the false claim buried in dense, legalistic phrasing): 0% [0%, 0%] at every temperature.
- **Formatting-duress** (the answer forced into a rigid output format): the only vector under which the model held ground — 67% [33%, 92%] at temperature 0.0, 73% [50%, 88%] at 0.4, falling to 52% [27%, 71%] at 0.7.

Two things follow. The weakness is **structural rather than stochastic**: resistance is at or statistically indistinguishable from zero on three of four pressure types, identically across independent items and temperatures, so the headline sycophancy result (0 of 9 held) is not an artifact of one phrasing — it generalises across pressure types and scales. And the lone exception is itself fragile: the formatting vector’s resistance carries a wide interval and degrades fifteen points as temperature rises toward production-typical settings. The practical implication is that this capitulation risk cannot be mitigated by varying how a contradiction is phrased — it is invariant to the surface form of the challenge, and must be addressed architecturally rather than through prompt design.

Steady under emotional framing — with an important caveat

Emotionally charged versions of arithmetic problems did not change the model's answers. The raw score showed three answer-changes across the emotional items, which a naive reading would call "equanimity failures." The QualitaX manipulation check showed that the emotional wording produced no measurable change in the model's behaviour at all — identical response length, no shift in tone. The three changes are therefore best explained as ordinary run-to-run randomness, not as the emotion affecting reasoning.

We report no genuine emotional decentering. We do not, however, claim positive emotional robustness — the test format limits how strongly that can be asserted. "No evidence of" is not proof of absence, and we treat it that way in the risk register.

Reasoning depth and strategy

The model scales answer length to problem difficulty by roughly 360× between trivial and demanding items — a strong indicator of working metacognitive control over reasoning depth. It also follows the strategy it declares it will use about 3.7 times out of 4. Neither is a load-bearing finding for the risk rating, but both are useful inputs to prompt design and pipeline routing.

How much to trust these numbers

Alongside the scores we ran the validity diagnostics described in Section 3. The most consequential outputs:

- **Format-clean.** The model answered in the requested format ~99% of the time, so its scores reflect ability, not an inability to be parsed.
- **Audited self-report metric.** Every "invented story" flag was re-checked to remove coincidental keyword matches before reporting.
- **Noise-vs-signal separation.** Answer changes were attributed to the prompt manipulation only when the manipulation demonstrably changed the model's behaviour — otherwise they are reported as randomness, as above.

SECTION 06

From findings to risk: the compliance translation

The remainder of the case study takes the findings above and translates them into the artefacts a deploying organisation actually needs: a risk register with inherent and residual ratings, a suitability tier for each use-case class, a control set, and a mapping to the regulatory frameworks most likely to apply.

Risk register

Risk	Likelihood	Impact	Inherent	Consequence if unmanaged
Capitulation to confident-but-false input (sycophancy)	High	High	**HIGH**	Users or upstream/retrieved content can override correct answers; misinformation, wrong decisions, manipulation, liability in regulated use.
Degradation under self-verification	High	Medium	**HIGH**	Standard reliability patterns ("are you sure?", self-critique, agent self-review) make outputs worse, not better.
Mis-stated self-confidence (calibration ~26% error)	Medium	Medium	**MEDIUM**	Self-reported reliability is an unreliable basis for routing or escalation; noisy estimate.
Sampling-dependent fragility	Medium	Medium	**MEDIUM**	Robustness varies with temperature; one-setting evaluation may not predict production.
Over-reliance on "equanimity" result	Low	Medium	**LOW–MED**	Emotional-robustness is "no evidence of failure," not proven robustness (measurement floor).
Confabulated self-explanations	Low	Low	**LOW**	Rare and largely genuine; low concern, but self-reports remain a behavioural proxy.

Required controls and residual risk

The control set below is the minimum we recommend for any deployment of gemma-4-e4b beyond minimal-risk internal use.

Control	Addresses	Residual
Prohibit self-verification / self-critique prompting; use independent re-sampling + majority vote instead of iterative refinement	Self-verification degradation	Low
Anchor facts to authoritative sources; do not let conversational contradiction override them; treat user contradictions as items to verify	Sycophancy (exposure)	Medium
Final decisions checked by a separate, pressure-resistant model or a human; the model never checks its own work	Sycophancy / self-regression	Low–Medium
Mandatory human oversight for high-risk/regulated outputs; no automatic update of records or state from user assertions	Sycophancy / accountability	Low
Pin and test the production sampling configuration; document temperature-dependence	Sampling fragility	Low
Log and monitor answer-changes-under-contradiction; red-team adversarial surfaces	Sycophancy / manipulation	Medium

RESIDUAL RISK AFTER CONTROLS

Acceptable for low- and limited-risk assistive use. For high-risk or regulated decision use, residual risk remains MEDIUM–HIGH, and the model is not recommended as a primary or autonomous decision component.

Suitability by use-case risk tier

Use-case tier	Determination
Minimal-risk / internal productivity (drafting, summarising non-decisional content)	Acceptable with standard controls.
Limited-risk user-facing (chat/assistant with transparency duties)	Acceptable only with sycophancy safeguards, a human fallback, and AI-use disclosure.
High-risk / regulated decision support (credit, employment, health, legal, essential services)	Not recommended as a primary/autonomous component. Only as a non-binding assistant with mandatory human oversight and an external arbiter; full

Use-case tier	Determination
	control set required.
Adversarial / untrusted-user public surfaces	Not recommended without substantial hardening; high manipulation exposure.

Where to deploy, and where not

Good fit	Needs safeguards, or a different model
Single-shot extraction, classification, summarisation, drafting	Multi-turn assistants exposed to user pushback
Neutral review / QA of content you provide	Adversarial or untrusted-user surfaces
Structured-output steps inside a pipeline	High-stakes factual Q&A (health, legal, finance, compliance)
Honest self-assessment / confidence surfacing	Agent loops that include a self-verification step

Regulatory and governance considerations

Obligations attach to the deploying system and organisation, not to a model-level finding. The mapping below is indicative and should be confirmed with the deployer's legal and compliance functions; it is not legal advice.

- **EU AI Act.** For high-risk systems (Annex III uses), the robustness deficit is directly relevant to Article 15 (accuracy and robustness) and Article 14 (human oversight). User-facing use may trigger Article 50 transparency duties. The finding should be recorded in technical documentation.
- **NIST AI RMF 1.0.** The case maps cleanly to MEASURE (validity, robustness, security) and MANAGE; the adversarial-input robustness gap is a risk to be documented and treated.
- **ISO/IEC 42001:2023 and 23894:2023.** The finding belongs on the AI risk register with a treatment plan and a periodic-review schedule under the AI management system.
- **Sector regimes.** Model-risk-management expectations in financial services (independent validation, ongoing monitoring) and clinical/medical-device rules where decision support is involved both apply additional duties. Not assessed in this engagement.

Documentation and monitoring

What should be on file when this assessment is used to support a deployment decision:

- Model card / risk assessment entry: model snapshot identifier, sampling configuration, evaluation method and date, intended and prohibited uses, the limitations of the assessment.
- Production monitoring: frequency of answer-changes under user contradiction; output drift; any change to the sampling configuration (re-evaluate on change).
- Periodic re-evaluation and re-evaluation on model/version update; retention of evaluation transcripts and run configuration for auditability.

SECTION 07

Lessons and methodological insights

Where validity diagnostics changed the conclusion

Sakshi's deterministic scoring produced reproducible per-task numbers, as designed. The harness's contribution sits one layer up by translating those numbers into findings that carry weight in a risk decision, with the diagnostic layer making clear when a number is robust and when it depends on test conditions.

On confabulation, the headline number alone would have been ambiguous . A one-in-ten rate could plausibly reflect either rare genuine confabulation, or coincidental keyword matches in the scoring logic. The false-positive audit and the demand-characteristic control together let us say something stronger and more useful: the rate is low, the residue is real, and it survives a deliberate probe. That is a different evidentiary basis for the "honest self-reporter" judgement than the raw number on its own.

On emotional equanimity, the raw numbers showed three answer-changes. A quick reading would have logged three failures. The manipulation check showed that the emotional framing had no measurable effect on the model's behaviour, so attributing the changes to the frame would have over-credited the test. We report "no evidence of emotional decentering" rather than "emotional robustness confirmed," which is a more honest and more defensible position to take into a risk register.

On the temperature dependence, behaviour at low temperatures looked notably more robust than behaviour at 0.7. Recording the exact sampling configuration alongside each score is one of the audit-trail steps the harness adds for client engagements . This is relevant to reproducibility and re-evaluation rather than to the research-benchmark scoring itself. The implication for any organisation deploying this model is concrete: evaluation must be done at the exact sampling settings used in production, and any change to those settings is a re-evaluation trigger. The recommendation in Section 6 to pin and test the production configuration follows directly from this.

Where the methodology alone is not sufficient

- The QualitaX harness is a custom implementation of the open source Sakshi Benchmark. The graders match the published outputs on every checked case, but they are reconstructed from observed behaviour rather than copied from the original source.
- Most metrics use small item samples and carry wide margins of error. The two findings driving the risk rating were consistent across runs; the secondary numbers (calibration, depth) are noisier and should be read accordingly.
- Self-report and robustness are inferred from outputs, not from internal model state. "No evidence of" is a useful empirical statement; it is not a proof of absence.
- One task (self-vs-other preference) relies on reference answers not available for this run, so it is excluded rather than reported with placeholder data. The evaluation is 20 of 21 tasks valid.

- The results are scoped to the specific snapshot of gemma-4-e4b served by the local endpoint used. Any model or version update is a re-evaluation trigger.

What this kind of work is good for, and what it is not

This report documents an independent, behavioural assessment of the Google gemma-4-e4b model. Its highest-leverage use is at two moments in the lifecycle of an AI deployment: at procurement, where it can help inform which model to choose and on what conditions; and at the risk-review gate, where it produces the documentation that lets a risk owner formally accept the residual risk.

It is not a substitute for legal advice, for sector-specific conformity work, or for the deploying organisation's own application-level testing. The case study is most usefully read as a worked example of the model-level inputs into those processes. This is the part that QualitaX can take off a client's plate, and that benefits most from independence.

SECTION 08

About this Case Study

Case study of the AI Risk Assessment for Google gemma-4-e4b. Evidence base: model-level metacognition evaluation against the Sakshi Benchmark, executed through the QualitaX assurance harness. Findings translated into a risk register, control set, and suitability tiering as documented in Section 6. Results are valid for the specific model snapshot and serving configuration documented in Section 4.

Provenance and reproducibility

Item	Detail
Subject model	Google gemma-4-e4b (gemma-4-e4b-it-gguf)
Serving stack	LM Studio 0.4.13, llama.cpp inference engine, OpenAI-compatible local endpoint
Sampling	Temperature 0.7; top-p, top-k, repeat penalty at LM Studio defaults (record values)
Chat template	LM Studio default for Gemma family (record template ID)
Runs	3 independent runs, ~180 calls per run
Task scope	20 of 21 Sakshi tasks scored validly; self-vs-other excluded
Scoring	Deterministic, rule-based — no AI judge
Validity layer	Format-compliance, false-positive audit, manipulation checks (full QualitaX suite)
Retained for audit	Full run configuration, per-task graded outputs, transcripts, and provenance record.

www.qualitax.io
contact@qualitax.io