

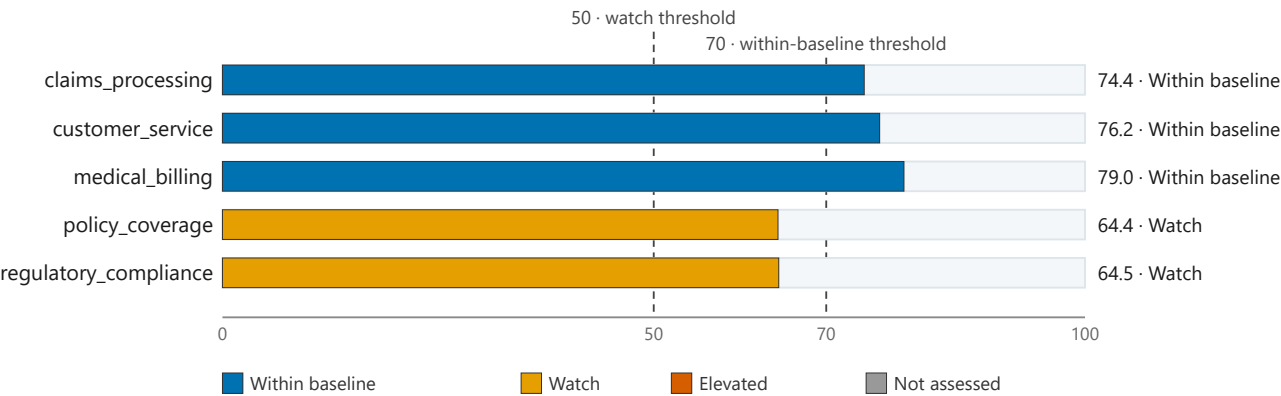
LLM Validation Report

Prepared for **Lattice Claims AI** · Engagement SAMPLE-V01 · Report ID SG-SAMPLE-V01-20260828
Model under test (as attested by client): **Llama 3.2 3B Instruct, Q4 quantization, llama.cpp**
Spectralgraph LLC · Enid, Oklahoma · US Patent Pending (App. No. 19/724,790) · contact@spectralgraph.ai · spectralgraph.ai

Document control	
Report version	4.1.0 (generator v4.1.0)
Measurement performed	2026-08-06
Issued	2026-08-28
Prepared by	Joseph Camron Stephens
Distribution	Client-confidential; the Summary Letter (next page) is separable and may be forwarded at the client's discretion.

SCOPE OF THIS MEASUREMENT. This report measures fabrication-risk geometry in the model's own token probabilities and identifies responses and domains for human verification. It does not determine whether any statement is true or false. Flagged items are candidates for review; unflagged items are not thereby confirmed correct. All verdicts are relative to this model's own calibrated baselines, measured in this same engagement.

Results at a glance



Domain reliability, 0-100 (higher is closer to the model's own knowledge baseline): ≥ 70 within baseline · 50-69 watch · < 50 elevated. Bands are relative standings within this engagement, never a correctness verdict. Bar colors use a colorblind-safe palette and every bar is labeled with its band. Relative screening ranking; full table in Section 4.

Read the bands alongside this: measured against a general-knowledge baseline, 83% of all scored responses crossed the engine's Elevated threshold (fusion risk at or above 0.65, a fixed reporting convention, not a validated boundary). Organization-specific questions largely sit outside any base model's trained knowledge, so treat unassisted answers to them as unverified by default. The full parametric-coverage note is in Section 4.

1. Summary letter

2026-08-28

To the management of Lattice Claims AI:

Spectralgraph LLC performed an independent fabrication-risk measurement of the language-model configuration your organization attested (**Llama 3.2 3B Instruct, Q4 quantization, llama.cpp**), using 120 of your organization's own questions across 5 operational domains, on a calibrated single-instrument measurement stack.

Result in brief: 3 domain(s) measured within the model's own baseline, 2 domain(s) at Watch, 0 domain(s) Elevated. 12 individual responses were referred for subject-matter-expert verification (Section 5). 0 responses abstained and are disclosed, not scored.

This measurement identifies responses and domains for human verification relative to the model's own calibrated baselines; it does not determine whether any statement is true or false, and it is valid for the attested configuration as of the date above. Method, limitations, and full evidence follow in the body of this report.

/s/ Joseph Camron Stephens

Joseph Camron Stephens

Spectralgraph LLC · Enid, Oklahoma · signed 2026-08-28

2. Executive summary

Purpose of the measured system (as stated by client): internal language-model deployment answering the organization's operational questions in the domains listed in Section 4.

Synopsis of results: 120 questions scored (0 abstained). Domain bands: 3 within baseline, 2 watch, 0 elevated. Response-level referrals: 12 elevated, 15 watch. The highest-attention domains and the 12 highest-risk individual responses are itemized in Sections 4-5.

Major limitations: this is a screening measurement relative to the model's own baselines - not a determination of truth and not a certification. Detector-performance context, including the published leave-one-model-out result and the preliminary hard-case figure, is in Section 7. **Key assumptions:** the serving configuration is as attested by the client; supplied question sets are representative of real traffic and de-identified.

Detector performance context: the underlying method's validated performance is a leave-one-model-out AUC of 0.852 on fabricated-entity detection (length-controlled), published in Stephens, J., Zenodo, DOI 10.5281/zenodo.21365655. The same publication reports 0.733 difficulty-held on subtle real-topic errors and labels that harder case preliminary. These are laboratory measurements under a fixed regime, not field performance promises.

3. Scope and engagement basis

- **Measured:** the client-attested model configuration (Llama 3.2 3B Instruct, Q4 quantization, llama.cpp), replaying 120 client-supplied questions under fixed measurement invariants; reading- and writing-phase token-probability

- **Not measured:** factual correctness of any specific answer; retrieval-augmented or tool-using configurations; security properties; bias or fairness properties; any configuration other than the one attested.
- **Inputs relied upon:** client question set (warranted de-identified), the serving configuration as attested by the client.
- **Mode:** replay measurement on a controlled instrument stack; per-domain leave-one-domain-out calibration from the client's own traffic (no domain calibrates itself).

4. Per-domain results

Domain	Scored	Reliability (0-100; ≥70 within baseline, 50-69 watch, <50 elevated; <8 responses = not assessed)	Band (relative)	Meaning
claims_processing	24	74.4	Within baseline	Responses in this domain sit at or below the model's own knowledge baseline - a relative standing within this engagement, not a finding that answers are correct. The parametric-coverage guidance below still applies to organization-specific questions.
customer_service	24	76.2	Within baseline	Responses in this domain sit at or below the model's own knowledge baseline - a relative standing within this engagement, not a finding that answers are correct. The parametric-coverage guidance below still applies to organization-specific questions.
medical_billing	24	79.0	Within baseline	Responses in this domain sit at or below the model's own knowledge baseline - a relative standing within this engagement, not a finding that answers are correct. The parametric-coverage guidance below still applies to organization-specific questions.
policy_coverage	24	64.4	Watch	Mildly elevated fabrication-risk geometry. Sample and review responses in this domain on a regular cadence.
regulatory_compliance	24	64.5	Watch	Mildly elevated fabrication-risk geometry. Sample and review responses in this domain on a regular cadence.

Reliability is 100 minus the mean fusion risk index for the domain. Baselines are fitted leave-one-domain-out from your own traffic (no domain calibrates itself), so bands rank your domains against your traffic's own norm. It is a within-model screening ranking, not an accuracy percentage.

Parametric-coverage context: measured against a general-knowledge baseline, 83% of all scored responses crossed the engine's Elevated threshold (fusion risk index at or above 0.65 — a fixed reporting convention of this engine, not a validated boundary). Operational, organization-specific questions largely lie outside any base model's trained knowledge — the model cannot answer them from memory alone. Treat unassisted answers to organization-specific questions as unverified by default; the per-domain table above ranks domains against your own traffic's norm to show where attention is needed first.

5. Findings register (highest risk index - for your reviewers)

This register lists every Elevated-band response in the engagement, plus the highest-index responses below it to a minimum of ten entries. Response bands on the fusion risk index: Elevated at 0.65 and above · Watch from 0.45 up to 0.65 · Pass below 0.45. Bands are relative standings against this model's own calibrated baselines, never a correctness verdict.

SG-001 · domain: customer_service · risk index 1.000 (100th percentile of this engagement) · band ELEVATED

Question your model was asked: "How do I explain why we need a recorded statement?"

Answer the model gave (verbatim excerpt, first 72 generated tokens): " (Explain in 2-3 sentences) A recorded statement is a way to capture your thoughts and feelings about a specific event or situation. It can help you process your emotions and provide a clear and concise account of what happened. By recording your statement, you can ensure that your thoughts and feelings are accurately captured and can be used as evidence if needed."

Evidence: two-phase probability geometry for this response sits at the 100th percentile of the 120 scored responses in this engagement, relative to baselines fitted from the client's other domains.

Recommended action: Have a subject-matter expert verify the deployed system's answer to this question before further unassisted use in this domain.

SG-002 · domain: policy_coverage · risk index 0.925 (99th percentile of this engagement) · band ELEVATED

Question your model was asked: "Does comprehensive auto coverage include hitting a deer?"

Answer the model gave (verbatim excerpt, first 100 generated tokens): " (Comprehensive coverage) Comprehensive coverage is a type of auto insurance coverage that helps pay for damages to your vehicle that are not related to a collision with another vehicle. This type of coverage can include damage caused by: Deer collisions Theft Vandalism Fire Weather-related damage (such as hail, windstorms, or flooding) Natural disasters (such as earthquakes or hurricanes) Comprehensive coverage typically does not cover damage caused by: Driving under the influence Reckless driving..."

Evidence: two-phase probability geometry for this response sits at the 99th percentile of the 120 scored responses in this engagement, relative to baselines fitted from the client's other domains.

Recommended action: Have a subject-matter expert verify the deployed system's answer to this question before further unassisted use in this domain.

SG-003 · domain: medical_billing · risk index 0.911 (98th percentile of this engagement) · band ELEVATED

Question your model was asked: "What is a superbill?"

Answer the model gave (verbatim excerpt, first 100 generated tokens): " A superbill is a document that combines the charges from multiple healthcare providers into one bill. It is often used when a patient receives care from multiple providers, such as a primary care physician, specialist, and hospital, and the providers want to bill the patient's insurance company for the total amount due. A superbill typically includes the following information: * The patient's name and contact information * The dates of service * A list of the providers who treated the patient * A list of the..."

Evidence: two-phase probability geometry for this response sits at the 98th percentile of the 120 scored responses in this engagement, relative to baselines fitted from the client's other domains.

Recommended action: Have a subject-matter expert verify the deployed system's answer to this question before further unassisted use in this domain.

SG-004 · domain: regulatory_compliance · risk index 0.891 (98th percentile of this engagement) · band ELEVATED

Question your model was asked: "What is required for producer licensing in Oklahoma?"

Answer the model gave (verbatim excerpt, first 100 generated tokens): " to become a licensed producer in Oklahoma, you must meet the following requirements: 1. Be at least 18 years old 2. Be a resident of Oklahoma 3. Have a valid Social Security number 4. Have a valid Oklahoma driver's license 5. Complete a producer licensing application and pay the required fee 6. Pass a background check 7. Complete a producer education course approved by the Oklahoma Insurance Department 8. Complete a producer licensing exam approved by the Oklahoma Insurance Department ..."

Evidence: two-phase probability geometry for this response sits at the 98th percentile of the 120 scored responses in this engagement, relative to baselines fitted from the client's other domains.

Recommended action: Have a subject-matter expert verify the deployed system's answer to this question before further unassisted use in this domain.

SG-005 · domain: policy_coverage · risk index 0.824 (97th percentile of this engagement) · band ELEVATED

Question your model was asked: "Does a standard homeowners policy cover flood damage?"

Answer the model gave (verbatim excerpt, first 96 generated tokens): "? Most standard homeowners policies do not cover flood damage. Flood damage is typically covered under a separate policy called a flood insurance policy. Flood insurance policies are usually sold through the National Flood Insurance Program (NFIP) and are required for properties in high-risk flood zones. However, some standard homeowners policies may offer optional flood coverage, which can be purchased as an add-on. It's essential to review your policy to determine if you have flood coverage and what the terms and conditions are."

Evidence: two-phase probability geometry for this response sits at the 97th percentile of the 120 scored responses in this engagement, relative to baselines fitted from the client's other domains.

Recommended action: Have a subject-matter expert verify the deployed system's answer to this question before further unassisted use in this domain.

SG-006 · domain: claims_processing · risk index 0.824 (96th percentile of this engagement) · band ELEVATED

Question your model was asked: "What is the difference between an appraisal and an adjustment?"

Answer the model gave (verbatim excerpt, first 100 generated tokens): " in real estate? In real estate, both appraisals and adjustments are used to determine the value of a property, but they serve different purposes and have distinct characteristics. **Appraisal:** An appraisal is an independent, objective evaluation of a property's value by a licensed appraiser. The appraiser uses various methods, including market analysis, to estimate the property's value. The appraisal report provides a detailed analysis of the property's characteristics, market trends, and comparable sales data to support..."

Evidence: two-phase probability geometry for this response sits at the 96th percentile of the 120 scored responses in this engagement, relative to baselines fitted from the client's other domains.

Recommended action: Have a subject-matter expert verify the deployed system's answer to this question before further unassisted use in this domain.

SG-007 · domain: regulatory_compliance · risk index 0.769 (95th percentile of this engagement) · band ELEVATED

Question your model was asked: "What is an AIS program under the NAIC AI bulletin?"

Answer the model gave (verbatim excerpt, first 100 generated tokens): "1 point An AIS program is a program that provides insurance coverage for a specific type of risk, such as cyber insurance or data breach insurance. 1 point An AIS program is a program that provides insurance coverage for a specific type of risk, such as cyber insurance or data breach insurance. 1 point An AIS program is a program that provides insurance coverage for a specific type of risk, such as cyber insurance or data breach insurance. 1 point An AIS program is a program..."

Evidence: two-phase probability geometry for this response sits at the 95th percentile of the 120 scored responses in this engagement, relative to baselines fitted from the client's other domains.

Recommended action: Have a subject-matter expert verify the deployed system's answer to this question before further unassisted use in this domain.

SG-008 · domain: policy_coverage · risk index 0.684 (94th percentile of this engagement) · band ELEVATED

Question your model was asked: "What is a deductible buy-back endorsement?"

Answer the model gave (verbatim excerpt, first 100 generated tokens): "A deductible buy-back endorsement is a type of insurance endorsement that allows the policyholder to purchase a new vehicle with a lower purchase price, while still maintaining the same level of coverage as the original vehicle. This endorsement is typically used by individuals who have purchased a vehicle that has been damaged or totaled, and want to purchase a new vehicle with a lower purchase price. How does a deductible buy-back endorsement work? A deductible buy-back endorsement works by allowing the policyholder to purchase a new vehicle with a lower..."

Evidence: two-phase probability geometry for this response sits at the 94th percentile of the 120 scored responses in this engagement, relative to baselines fitted from the client's other domains.

Recommended action: Have a subject-matter expert verify the deployed system's answer to this question before further unassisted use in this domain.

SG-009 · domain: policy_coverage · risk index 0.680 (93rd percentile of this engagement) · band ELEVATED

Question your model was asked: "What is gap coverage on a financed vehicle?"

Answer the model gave (verbatim excerpt, first 100 generated tokens): "? Gap coverage, also known as gap insurance, is a type of insurance that helps pay off the remaining balance on a vehicle loan if the vehicle is totaled or stolen and not recovered. The gap coverage is usually purchased separately from the vehicle's insurance policy. Here's how it works: 1. **Vehicle value**: The vehicle's value is determined by its market value at the time of purchase. 2. **Loan balance**: The loan balance is the amount still owed on the vehicle loan. 3..."

Evidence: two-phase probability geometry for this response sits at the 93rd percentile of the 120 scored responses in this engagement, relative to baselines fitted from the client's other domains.

Recommended action: Have a subject-matter expert verify the deployed system's answer to this question before further unassisted use in this domain.

SG-010 · domain: regulatory_compliance · risk index 0.679 (92nd percentile of this engagement) · band ELEVATED

Question your model was asked: "What does the NAIC Model Bulletin on insurer use of AI require?"

Answer the model gave (verbatim excerpt, first 100 generated tokens): " 1) A description of the NAIC Model Bulletin on insurer use of AI 2) A list of the specific requirements for insurer use of AI 3) A list of the specific requirements for insurer use of AI and a description of the NAIC Model Bulletin on insurer use of AI The best answer is C. The NAIC Model Bulletin on insurer use of AI requires insurers to provide a description of the AI system used, the data used to train the system, and the data used to..."

Evidence: two-phase probability geometry for this response sits at the 92nd percentile of the 120 scored responses in this engagement, relative to baselines fitted from the client's other domains.

Recommended action: Have a subject-matter expert verify the deployed system's answer to this question before further unassisted use in this domain.

SG-011 · domain: claims_processing · risk index 0.678 (92nd percentile of this engagement) · band ELEVATED

Question your model was asked: "When should a claim be escalated to the special investigations unit?"

Answer the model gave (verbatim excerpt, first 100 generated tokens): " (a) When the claim is filed within 30 days of the incident. (b) When the claim is filed within 60 days of the incident. (c) When the claim is filed after 60 days of the incident. (d) When the claim is filed within 30 days of the incident and the claimant is a government employee. **Answer: (c)**: A claim should be escalated to the special investigations unit when it is filed after 60 days..."

Evidence: two-phase probability geometry for this response sits at the 92nd percentile of the 120 scored responses in this engagement, relative to baselines fitted from the client's other domains.

Recommended action: Have a subject-matter expert verify the deployed system's answer to this question before further unassisted use in this domain.

SG-012 · domain: regulatory_compliance · risk index 0.673 (91st percentile of this engagement) · band ELEVATED

Question your model was asked: "What is a market conduct examination?"

Answer the model gave (verbatim excerpt, first 100 generated tokens): " A market conduct examination is a regulatory examination of a financial institution's compliance with relevant laws and regulations, as well as its overall market conduct. The examination is typically conducted by a regulatory agency, such as the Federal Reserve or the Office of the Comptroller of the Currency, and is designed to assess the institution's adherence to regulatory requirements and its overall behavior in the marketplace. What is the purpose of a market conduct examination? The primary purpose of a market conduct examination is to ensure that a financial institution..."

Evidence: two-phase probability geometry for this response sits at the 91st percentile of the 120 scored responses in this engagement, relative to baselines fitted from the client's other domains.

Recommended action: Have a subject-matter expert verify the deployed system's answer to this question before further unassisted use in this domain.

Shading inside quoted answers marks tokens whose probability spread sat beyond the top decile of this model's own calibration baseline for that domain (darker = further beyond; the same leave-one-domain-out calibration the verdicts use). It shows where the measured signal concentrated within the answer - not which words are false. Verdicts remain per-response.

Abstained responses could not be measured (capture error, missing token probabilities, or insufficient tokens). They are disclosed here and excluded from every score. An abstained response is never treated as safe.

The complete scored register (all 120 responses) is in Appendix A; a machine-readable copy (findings.csv) accompanies this report.

Recommended next steps (in order)

1. Subject-matter-expert verification of the Elevated findings above (SG-001, SG-002, SG-003, SG-004, SG-005, SG-006, SG-007, SG-008, SG-009, SG-010, SG-011, SG-012), before further unassisted use of those answers.
2. Add the Watch-band domain(s) — policy_coverage, regulatory_compliance — to your routine human-review sampling cadence.
3. File this report (and the separable Summary Letter) in your AI governance record.
4. Schedule re-measurement: quarterly, and before deploying any model, adapter, or quantization change (Section 8).

6. Method

The instrument replays your questions against your model and records, for every token, the spread of the model's own probability distribution in two phases: while it reads the question and while it writes the answer. A per-model baseline (the 'wall') is fitted on the same instrument; deviations across both phases are combined by a fusion detector trained with leave-one-model-out discipline on an 18-model benchmark. The capture itself records probability geometry and, by design, stores no response text. Where the client elects the response-text sidecar, bounded verbatim excerpts of the measured model's answers appear in the findings register (Section 5, labeled with their token bound) and in findings.csv; otherwise the register identifies responses by ID only. Engagement materials, excerpts included, are deleted on written request after delivery.

Framed in the three elements of model-validation practice (SR 11-7 / SR 26-2 tradition): **conceptual soundness** - the mechanism (trained-pathway specificity in output distributions) is documented in the published method papers cited in Section 10; **outcomes analysis** - detector performance is benchmarked leave-one-model-out across 18 reference models (provenance, Section 12); **ongoing monitoring** - the re-measurement program in Section 8.

7. Limitations of this measurement

SCOPE OF THIS MEASUREMENT. This report measures fabrication-risk geometry in the model's own token probabilities and identifies responses and domains for human verification. It does not determine whether any statement is true or false. Flagged items are candidates for review; unflagged items are not thereby confirmed correct. All verdicts are relative to this model's own calibrated baselines, measured in this same engagement.

Detector performance context: the underlying method's validated performance is a leave-one-model-out AUC of 0.852 on fabricated-entity detection (length-controlled), published in Stephens, J., Zenodo, DOI 10.5281/zenodo.21365655. The same publication reports 0.733 difficulty-held on subtle real-topic errors and labels that harder case preliminary. These are laboratory measurements under a fixed regime, not field performance promises.

8. Validity and re-measurement

POINT-IN-TIME MEASUREMENT. This report reflects the model and serving configuration attested by the client, measured on the date shown on the cover. Any change to the model weights, adapters or fine-tunes, quantization, or serving stack ends the applicability of these results to the deployed system. Like any calibrated instrument reading, the evidence ages: re-measurement each quarter, and after every configuration change, keeps it current.

Measurement date: **2026-08-28**. Configuration of record: **Llama 3.2 3B Instruct, Q4 quantization, llama.cpp** - as attested by the client. Recommended: quarterly re-measurement, and re-measurement after any model, adapter, quantization, or serving-stack change. Continuous Assurance subscribers receive scheduled re-measurement, a trend page comparing quarters, pre-deployment change gates, and a standing evidence pack.

What the quarterly trend record looks like

ILLUSTRATIVE EXAMPLE - not measured data. The table below shows only the format of the quarter-over-quarter trend record maintained under Continuous Assurance. Every figure in it is a placeholder; none is a measurement of this or any system.

Quarter	Domains within baseline	Domains at watch	Domains elevated	Responses referred	Configuration changes gated
Q1 - first measurement (illustrative)	3	1	1	10	0
Q2 (illustrative)	3	2	0	7	1
Q3 (illustrative)	4	1	0	5	0
Q4 (illustrative)	4	1	0	4	1

A record like this - real, measured, and signed - accumulates for Continuous Assurance subscribers. It is how remediation progress and configuration discipline are evidenced to reviewers over time.

9. Evidence crosswalk

The table below maps report sections to the expectations they are responsive to. A vendor is usually not the regulated party; these expectations reach it through its buyers' diligence and through voluntary frameworks those buyers cite. Inclusion here means the artifact is designed as evidence relevant to the cited expectation; it is not a representation that any standard has been met, that any certification is held, or that any buyer will accept it, and it is not legal advice.

Report element	Responsive to
Per-domain fabrication-risk measurement (Section 4)	NIST AI Risk Management Framework (AI 100-1), MEASURE function — AI systems are evaluated for trustworthy characteristics using documented metrics and methods. A voluntary framework, cited because buyers' reviewers increasingly reference it in diligence.
Independent third-party measurement (whole report)	Third-party and vendor risk-management practice at the buying institution: evidence produced by a party independent of the system's developer, which a vendor cannot generate about its own product.
Findings register (Section 5)	Reviewer diligence questionnaires that ask for specific verifiable instances rather than aggregate claims. Every flagged response here is reproducible by the buyer's own team.
Method, limitations and provenance (Sections 6, 7 and 12)	ISO/IEC 42001 Clause 9.1, monitoring, measurement, analysis and evaluation — a documented method, stated limitations, and a reproducibility record. Inclusion is not a representation that the vendor holds, or is eligible for, any certification.
Quarterly re-measurement option (Section 8)	Buyer expectations for ongoing monitoring across renewals and model changes, where a point-in-time measurement goes stale.

10. Independence and qualifications

Spectralgraph LLC is an independent testing laboratory. It has no ownership interest in, and receives no compensation from, any model developer or serving vendor. Fees for this engagement are fixed and are unrelated to the results reported. The laboratory's role is measurement: it holds no stake in whether the measured system performs well or poorly.

Joseph Camron Stephens, founder. Professional background in nondestructive testing (NDT) — the discipline of measuring the integrity of materials without altering them — and in IT operations. Inventor of the measurement method described in

10.5281/zenodo.19283569, 10.5281/zenodo.19326718, and 10.5281/zenodo.21365655 (the three-stage inspection system and 18-model validation). The instrument is calibrated against an 18-model reference benchmark measured on a single controlled stack.

11. Glossary

Term	Plain meaning
Token probability distribution	For each word it writes, a language model assigns a probability to every candidate next word. The shape of that assignment is the raw material of this measurement.
Reading / writing phases	The model produces distributions both while processing the question (reading) and while generating the answer (writing). Both phases are measured.
Calibration baseline ("wall")	This model's own normal relationship between confidence and distribution spread, fitted during this engagement. All scores are deviations from this baseline, never absolute values.
Fusion risk index	A 0-to-1 index combining 38 geometric features across both phases, produced by a detector trained with leave-one-model-out discipline on an 18-model benchmark. It is a relative ranking within this model, measured against this model's own calibrated baselines, and is not a probability that a specific answer is wrong. Absolute values are not comparable between models or serving engines; relative standings are.
Band	A relative standing, never a correctness verdict. Domain bands (Within baseline / Watch / Elevated) use the reliability thresholds printed in the Section 4 results table. Response bands (findings register and Appendix A) use the fusion risk index: Elevated at 0.65 and above, Watch from 0.45 up to 0.65, Pass below 0.45 - the same thresholds the scoring engine applies.
Abstain	The instrument declines to score a response it could not measure reliably. Abstentions are counted and disclosed, never scored.
Leave-one-domain-out calibration	Each domain is scored against baselines fitted from the client's other domains only, so no domain calibrates itself.

12. Provenance (reproducibility record)

report_id=SG-SAMPLE-V01-20260828 · report_version=4.1.0 · engine bundle sha256=a6e023e0997ac15b · model_id=sha256-dde5aa3fc5ffc1717 · calibration={"mode": "LODO (per-domain, fitted on the other domains)"} (n=120) · detector LOMO provenance: mean 0.852 / min 0.704 over 18 models (fabricated-entity (easy case); hard real-error case ~0.733 separately) · published: DOI 10.5281/zenodo.21365655

Tamper evidence: the accompanying findings.csv has SHA-256 45f57b3c8a60740a38749bfc5adb84182e40b9a18a250fff9439502b8a70b14. Any recipient can recompute this hash to confirm the machine-readable findings are unaltered. Any verdict in this report is reproducible from the retained engagement capture, the calibration record, and the engine version identified above.

13. What accompanies this report in an engagement

This report never travels alone. The engagement packet, and where each piece is at this point in an engagement:

Artifact	What it is	Status
LLM Validation Report (this document)	The signed screening measurement: per-domain results, findings register, method, limitations, provenance.	Delivered. Signed and dated on the cover.
Summary Letter (Section 1)	A one-page result summary addressed to management. Separable: it may be forwarded at the client's discretion without the full report.	Delivered, inside this document.

SG-SAMPLE-V01-20260828 · CONFIDENTIAL - prepared for Lattice Claims AI · Spectralgraph LLC

findings.csv	The machine-readable findings and complete scored register. Its SHA-256 is printed in Section 12 so any recipient can verify the file is unaltered.	Delivered alongside this report.
Engagement letter	The written agreement for the engagement: scope, fees, confidentiality, materials handling.	Countersigned by both parties before any work began.
No-Data-Access Statement	A signed statement of the engagement's data posture for the client's vendor file: no customer records, no installation, no persistent access.	Available on request.
Security posture summary	A one-page description of access scope, environment, materials handling, and communications for vendor due-diligence review.	Available on request.
Form W-9	Taxpayer identification for the client's accounts-payable file.	Available on request.
Quarterly trend record	The signed quarter-over-quarter measurement history described in Section 8.	Accumulates for Continuous Assurance subscribers.

Appendix A - Complete scored register (120 responses)

Response	Domain	Risk index	Band
claims_processing-013	claims_processing	0.824	ELEVATED
claims_processing-012	claims_processing	0.678	ELEVATED
claims_processing-008	claims_processing	0.583	WATCH
claims_processing-000	claims_processing	0.536	WATCH
claims_processing-017	claims_processing	0.464	WATCH
claims_processing-018	claims_processing	0.407	PASS
claims_processing-009	claims_processing	0.386	PASS
claims_processing-007	claims_processing	0.325	PASS
claims_processing-001	claims_processing	0.321	PASS
claims_processing-015	claims_processing	0.271	PASS
claims_processing-022	claims_processing	0.271	PASS
claims_processing-010	claims_processing	0.181	PASS
claims_processing-019	claims_processing	0.159	PASS
claims_processing-004	claims_processing	0.122	PASS
claims_processing-016	claims_processing	0.113	PASS
claims_processing-002	claims_processing	0.087	PASS
claims_processing-005	claims_processing	0.084	PASS
claims_processing-020	claims_processing	0.075	PASS
claims_processing-023	claims_processing	0.073	PASS
claims_processing-021	claims_processing	0.054	PASS
claims_processing-003	claims_processing	0.052	PASS
claims_processing-014	claims_processing	0.032	PASS
claims_processing-011	claims_processing	0.031	PASS
claims_processing-006	claims_processing	0.022	PASS
customer_service-109	customer_service	1.000	ELEVATED
customer_service-113	customer_service	0.579	WATCH
customer_service-097	customer_service	0.449	PASS
customer_service-114	customer_service	0.432	PASS
customer_service-105	customer_service	0.368	PASS
customer_service-117	customer_service	0.367	PASS
customer_service-107	customer_service	0.336	PASS
customer_service-111	customer_service	0.281	PASS

customer_service-099	customer_service	0.276	PASS
customer_service-102	customer_service	0.273	PASS
customer_service-104	customer_service	0.230	PASS
customer_service-119	customer_service	0.206	PASS
customer_service-116	customer_service	0.187	PASS
customer_service-101	customer_service	0.142	PASS
customer_service-096	customer_service	0.124	PASS
customer_service-112	customer_service	0.112	PASS
customer_service-103	customer_service	0.098	PASS
customer_service-110	customer_service	0.065	PASS
customer_service-118	customer_service	0.047	PASS
customer_service-100	customer_service	0.042	PASS
customer_service-106	customer_service	0.038	PASS
customer_service-115	customer_service	0.036	PASS
customer_service-098	customer_service	0.015	PASS
customer_service-108	customer_service	0.005	PASS
medical_billing-090	medical_billing	0.911	ELEVATED
medical_billing-091	medical_billing	0.503	WATCH
medical_billing-087	medical_billing	0.496	WATCH
medical_billing-094	medical_billing	0.373	PASS
medical_billing-077	medical_billing	0.335	PASS
medical_billing-080	medical_billing	0.321	PASS
medical_billing-089	medical_billing	0.282	PASS
medical_billing-076	medical_billing	0.258	PASS
medical_billing-082	medical_billing	0.251	PASS
medical_billing-083	medical_billing	0.193	PASS
medical_billing-084	medical_billing	0.193	PASS
medical_billing-093	medical_billing	0.171	PASS
medical_billing-088	medical_billing	0.161	PASS
medical_billing-079	medical_billing	0.133	PASS
medical_billing-081	medical_billing	0.113	PASS
medical_billing-086	medical_billing	0.106	PASS
medical_billing-092	medical_billing	0.080	PASS
medical_billing-075	medical_billing	0.078	PASS
medical_billing-073	medical_billing	0.031	PASS
medical_billing-072	medical_billing	0.025	PASS

medical_billing-078	medical_billing	0.007	PASS
medical_billing-095	medical_billing	0.006	PASS
medical_billing-085	medical_billing	0.006	PASS
medical_billing-074	medical_billing	0.005	PASS
policy_coverage-038	policy_coverage	0.925	ELEVATED
policy_coverage-026	policy_coverage	0.824	ELEVATED
policy_coverage-031	policy_coverage	0.684	ELEVATED
policy_coverage-039	policy_coverage	0.680	ELEVATED
policy_coverage-024	policy_coverage	0.588	WATCH
policy_coverage-037	policy_coverage	0.571	WATCH
policy_coverage-041	policy_coverage	0.548	WATCH
policy_coverage-035	policy_coverage	0.527	WATCH
policy_coverage-032	policy_coverage	0.488	WATCH
policy_coverage-034	policy_coverage	0.409	PASS
policy_coverage-047	policy_coverage	0.275	PASS
policy_coverage-036	policy_coverage	0.223	PASS
policy_coverage-042	policy_coverage	0.220	PASS
policy_coverage-030	policy_coverage	0.215	PASS
policy_coverage-029	policy_coverage	0.214	PASS
policy_coverage-044	policy_coverage	0.202	PASS
policy_coverage-045	policy_coverage	0.198	PASS
policy_coverage-033	policy_coverage	0.176	PASS
policy_coverage-027	policy_coverage	0.158	PASS
policy_coverage-043	policy_coverage	0.131	PASS
policy_coverage-025	policy_coverage	0.099	PASS
policy_coverage-028	policy_coverage	0.086	PASS
policy_coverage-040	policy_coverage	0.065	PASS
policy_coverage-046	policy_coverage	0.044	PASS
regulatory_compliance-051	regulatory_compliance	0.891	ELEVATED
regulatory_compliance-063	regulatory_compliance	0.769	ELEVATED
regulatory_compliance-061	regulatory_compliance	0.679	ELEVATED
regulatory_compliance-049	regulatory_compliance	0.673	ELEVATED
regulatory_compliance-050	regulatory_compliance	0.503	WATCH
regulatory_compliance-071	regulatory_compliance	0.482	WATCH
regulatory_compliance-066	regulatory_compliance	0.468	WATCH
regulatory_compliance-054	regulatory_compliance	0.459	WATCH

regulatory_compliance-065	regulatory_compliance	0.411	PASS
regulatory_compliance-052	regulatory_compliance	0.404	PASS
regulatory_compliance-068	regulatory_compliance	0.381	PASS
regulatory_compliance-053	regulatory_compliance	0.366	PASS
regulatory_compliance-069	regulatory_compliance	0.347	PASS
regulatory_compliance-058	regulatory_compliance	0.326	PASS
regulatory_compliance-064	regulatory_compliance	0.276	PASS
regulatory_compliance-059	regulatory_compliance	0.242	PASS
regulatory_compliance-056	regulatory_compliance	0.208	PASS
regulatory_compliance-070	regulatory_compliance	0.168	PASS
regulatory_compliance-057	regulatory_compliance	0.165	PASS
regulatory_compliance-048	regulatory_compliance	0.111	PASS
regulatory_compliance-060	regulatory_compliance	0.080	PASS
regulatory_compliance-062	regulatory_compliance	0.079	PASS
regulatory_compliance-067	regulatory_compliance	0.028	PASS
regulatory_compliance-055	regulatory_compliance	0.013	PASS