

LLM Validation Report

Prepared for **Meridian Mutual Insurance (SAMPLE - fictional client, real measurements)** · Engagement SAMPLE-001 · Report ID SG-SAMPLE-001-20260703

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	3.0.0 (generator v3.0.0)
Measurement performed	2026-07-03
Issued	2026-07-03
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

3 Domains within baseline	2 Domains at watch	0 Domains elevated	10 Responses referred for review	0 Abstained (disclosed, not scored)
------------------------------	-----------------------	-----------------------	-------------------------------------	--

claims_processing	74.4	
customer_service	76.2	
medical_billing	79.0	
policy_coverage	64.4	
regulatory_compliance	64.5	

Domain reliability, 0-100 (higher is closer to the model's own knowledge baseline): ≥70 within baseline · 50-69 watch · <50 elevated. PROVISIONAL screening ranking; full table in Section 4.

1. Summary letter

2026-07-03

To the management of Meridian Mutual Insurance (SAMPLE - fictional client, real measurements):

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. 10 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.

Joseph Camron Stephens
Spectralgraph LLC · Enid, Oklahoma

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 10 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, not a certification, and PROVISIONAL in its detector-performance figures (Section 6). **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 approximately 0.85 on fabricated-entity detection (length-controlled, PROVISIONAL pending larger benchmarks) and 0.733 difficulty-held on subtle real-topic errors. 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 geometry.

- **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	Meaning
claims_processing	24	74.4	Within baseline	Responses in this domain sit at or below the model's own knowledge baseline. Routine spot-checks are sufficient.
customer_service	24	76.2	Within baseline	Responses in this domain sit at or below the model's own knowledge baseline. Routine spot-checks are sufficient.
medical_billing	24	79.0	Within baseline	Responses in this domain sit at or below the model's own knowledge baseline. Routine spot-checks are sufficient.
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 PROVISIONAL 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 showed geometry resembling this model's fabrication baseline. 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)

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?

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?

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?

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?

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?

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?

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?

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?

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 (93th percentile of this engagement) · band ELEVATED

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

Evidence: two-phase probability geometry for this response sits at the 93th 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 (92th percentile of this engagement) · band ELEVATED

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

Evidence: two-phase probability geometry for this response sits at the 92th 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.

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), 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

one-model-out discipline on an 18-model benchmark. No client content is retained beyond the engagement; the capture records probability geometry.

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 9; **outcomes analysis** - detector performance is benchmarked leave-one-model-out across 18 reference models (provenance, Section 10); **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 approximately 0.85 on fabricated-entity detection (length-controlled, PROVISIONAL pending larger benchmarks) and 0.733 difficulty-held on subtle real-topic errors. 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: **2026-07-03**. 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.

9. Regulatory crosswalk

The table below maps report sections to the regulatory language they are responsive to. Inclusion here means the artifact is designed as evidence relevant to the cited expectation; it is not a representation that any regulatory standard has been met, and it is not legal advice.

Report element	Responsive to
Per-domain fabrication-risk measurement (Section 4)	NAIC Model Bulletin: Use of AI Systems by Insurers — "documented pre- and post-deployment testing" of AI systems (adopted in Oklahoma as OID Bulletin No. 2024-11)
Independent third-party measurement (whole report)	NAIC bulletin AIS Program expectations for verification and testing independent of the system's developer
Quarterly re-measurement option (Section 8)	NAIC bulletin expectations for ongoing monitoring of AI systems in production

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 this report (US Patent Pending, App. No. 19/724,790). The method is published: Stephens, J., Zenodo, DOI 10.5281/zenodo.19283569 and DOI 10.5281/zenodo.19326718. 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. A within-model ranking signal, labeled PROVISIONAL; not a probability that a specific answer is wrong.
Band	Within baseline / Watch / Elevated - thresholds are printed in the results table.
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-001-20260703 · report_version=3.0.0 · 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) · status PROVISIONAL

Tamper evidence: the accompanying findings.csv has SHA-256 83e42e11e6409415c43a5ad5602fce6a3ecdb19a612542715392aaf1168dc3a1. 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.

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