International Journal of Technology and Applied Science (IJTAS)

E-ISSN: 2230-9004 ● Website: www.ijtas.com ● Email: editor@ijtas.com

Measuring Trust in Generative AI: A Framework for Accuracy, Bias, and Safety Metrics

Rajesh Lingam[1], Meena Masoodi[2], Pavan M[3]

[1]Independent Researcher, Boston, USA. lingamrajesh06@gmail.com
[2]Independent Researcher, USA. meena.masoodi17@gmail.com
[3]Independent Researcher, Bangalore, India. marampavankuma@gmail.com

Abstract:
Trust in Generative AI systems depends on demonstrable, measurable evidence that these systems behave accurately, fairly, and safely across diverse user populations and use cases. Despite growing industry and regulatory interest in trustworthy AI, practical frameworks for operationalizing trust measurement in production systems remain underdeveloped. Existing benchmarks such as HELM and BIG-Bench evaluate capability breadth but do not produce actionable trust scores suitable for continuous production monitoring. This paper proposes the **Trust Measurement Framework (TMF)** for Generative AI, defining quantifiable metrics across three dimensions: *accuracy* (semantic correctness, factual grounding, and citation precision), *bias* (demographic parity delta, equalized odds difference, and counterfactual sensitivity), and *safety* (harmful content rate, boundary robustness index, and adversarial resistance score). We describe metric definitions, measurement methodologies, and a weighted aggregation strategy that produces a single composite Trust Score for production LLM systems. Validated across three successive model versions in an enterprise document intelligence platform serving over ten million users, the TMF demonstrated stable predictive validity for user satisfaction outcomes (Pearson $r = 0.81$ across three model versions) and correctly detected a safety regression in a controlled pre-deployment scenario that accuracy-only evaluation would have missed. We discuss metric selection trade-offs, calibration approaches, and integration patterns with automated evaluation pipelines. The TMF provides a practical, reproducible foundation for organizations that must demonstrate trustworthy AI behavior to regulators, customers, and internal governance bodies.
Trustworthy AI, generative AI evaluation, large language models, bias measurement, AI safety, trust scoring, enterprise AI, responsible AI.

Introduction

deployment of Large Language Model (LLM)-powered products in enterprise settings has accelerated rapidly since the publication of GPT-3 and subsequent instruction-tuned variants . Products built on these foundations now serve tens of millions of users in high-stakes domains including legal document processing, healthcare information retrieval, financial document analysis, and enterprise knowledge management. As these products move from prototype to production, the question of *trustworthiness*—

whether users, enterprises, and regulators can rely on the AI system to behave as claimed—has become critical.

Trust in AI systems is not monolithic. A system may be highly accurate in aggregate while producing systematically biased outputs for minority demographic groups. It may behave safely under typical inputs while remaining vulnerable to adversarial manipulation. It may ground responses in source documents accurately while exhibiting harmful content generation tendencies in edge cases. Existing evaluation frameworks tend to measure these properties in isolation, using academic benchmarks designed for research comparisons rather than for ongoing production monitoring.

This gap creates a practical problem for organizations deploying generative AI in enterprise products: they lack a principled, quantitative mechanism to answer the question, *"Can we trust this model version in production?"* and to track the answer over time across successive model updates. The absence of such measurement also impedes regulatory compliance, as frameworks such as the NIST AI Risk Management Framework (AI RMF) and the EU AI Act explicitly require demonstrable evidence of trustworthy behavior but do not specify measurement methodologies.

This paper makes the following contributions:

1. We propose the **Trust Measurement Framework (TMF)**, a three-dimensional framework defining nine quantifiable metrics for accuracy, bias, and safety in production generative AI systems.
2. We describe measurement methodologies for each metric, including dataset construction, evaluation protocols, and computation procedures reproducible by practitioners.
3. We introduce a **weighted aggregation strategy** that produces a single composite Trust Score suitable for release-gate decisions, longitudinal tracking, and governance reporting.
4. We present empirical validation of the TMF across three model versions in a production enterprise document AI system, demonstrating predictive validity for user satisfaction and regression detection capability.
5. We discuss integration patterns for embedding TMF measurement into automated CI/CD evaluation pipelines.

The remainder of this paper is organized as follows. Section 2 reviews related work. Section 3 defines the TMF dimensions, metrics, and aggregation strategy. Section 4 describes measurement methodology. Section 5 presents experimental validation. Section 6 discusses trade-offs, calibration, and CI/CD integration. Section 7 concludes.

Background and Related Work

Language Model Evaluation Benchmarks

The measurement of language model quality has been dominated by task-specific benchmarks. GLUE and its successor SuperGLUE standardized evaluation on natural language understanding tasks. BIG-Bench extended this to 204 diverse tasks contributed by over 400 researchers, revealing that model performance improves with scale but remains poor in absolute terms on reasoning-heavy tasks.

The most comprehensive evaluation framework to date is HELM , which evaluates 30 language models across 42 scenarios using seven metrics including accuracy, calibration, robustness, fairness, bias, toxicity, and efficiency. HELM represents a significant advance in evaluation breadth. However, it is

designed for comparative research evaluation of model capabilities rather than for continuous production monitoring of a specific deployed system. Its computational cost and the requirement for ground-truth annotation make it impractical as a release-gate tool.

TruthfulQA specifically targets whether models generate truthful answers even when trained on human text containing falsehoods. Lin *et al.* found that larger models were less truthful, an important finding for enterprise deployments where scaling is a common quality improvement strategy. Ji *et al.* provide a comprehensive survey of hallucination in natural language generation, categorizing hallucinations as *intrinsic* (contradicting source material) or *extrinsic* (introducing information not in the source), directly informing our accuracy metric design.

Fairness and Bias Measurement

Algorithmic fairness has been studied extensively in supervised classification settings, with well-established definitions including demographic parity, equalized odds, and counterfactual fairness . The AI Fairness 360 toolkit provides implementations of over 70 fairness metrics for classification tasks.

For language models specifically, StereoSet measures stereotypical bias across gender, profession, race, and religion using a fill-in-the-blank paradigm. CrowS-Pairs provides 1,508 sentence pairs testing nine types of social bias in masked language models. WinoBias targets gender bias in coreference resolution, demonstrating that gendered occupational stereotypes significantly affect model behavior.

Safety in Generative AI

Gehman *et al.* introduced RealToxicityPrompts, a dataset of 100,000 naturally-occurring prompts demonstrating that language models can generate toxic content even from innocuous prompts. LaMDA demonstrates an industry approach to dialog safety through fine-tuning and external knowledge retrieval, explicitly noting that model scaling alone does not address safety. Bai *et al.* introduced Constitutional AI, a self-supervised approach to harmlessness using natural language principles. Ouyang *et al.* demonstrated that RLHF significantly reduces harmful outputs, with 1.3B parameter InstructGPT outputs preferred over 175B parameter GPT-3 outputs in human evaluations.

Responsible AI Governance Frameworks

Mitchell *et al.* proposed Model Cards, standardized documentation for AI systems detailing performance characteristics across demographic groups. The NIST AI RMF , published in January 2023, provides a voluntary framework with four core functions: GOVERN, MAP, MEASURE, and MANAGE. The MEASURE function calls for quantitative metrics across trustworthiness dimensions but leaves metric specification to implementing organizations—the gap the TMF directly addresses.

Gap Analysis

Reviewing existing work, three gaps motivate the TMF: (1) **Production monitoring gap:** Existing benchmarks are designed for research comparisons, not continuous production monitoring. (2) **Generative adaptation gap:** Bias and safety metrics from classification settings require adaptation for generative outputs. (3) **Integration gap:** No existing work addresses how trust metrics should be aggregated, calibrated, and integrated into CI/CD pipelines.

Trust Measurement Framework

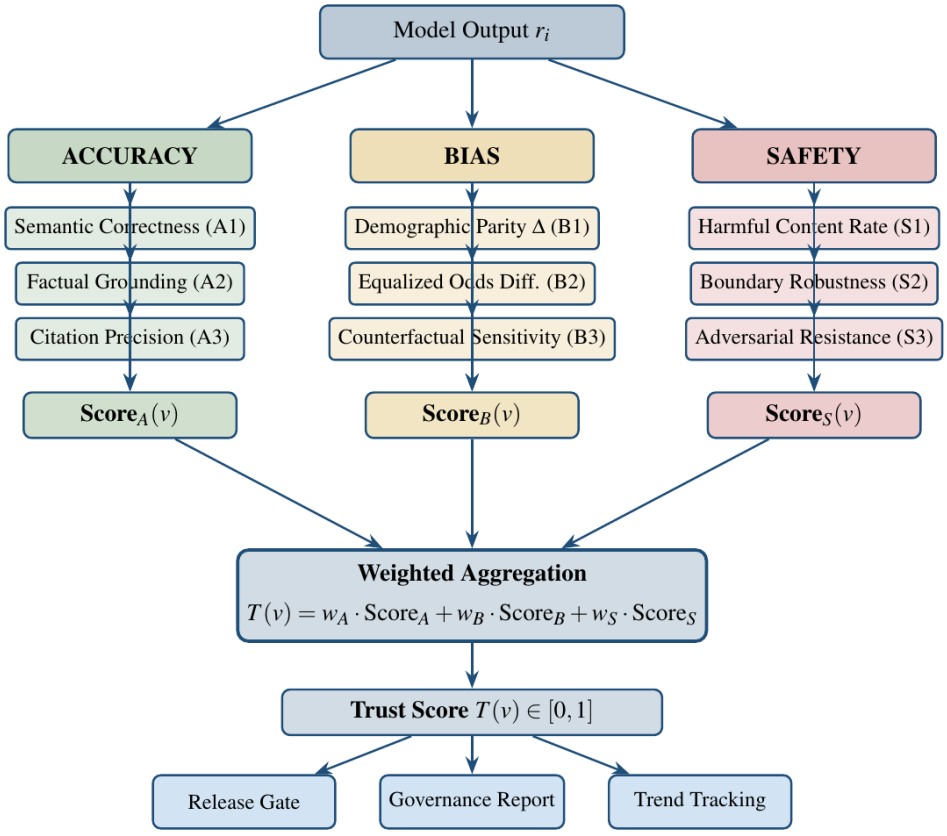
Design Principles

The TMF is designed around four principles:

- *Quantifiability*: Every dimension produces numerical scores in [0,1] to enable comparison and aggregation.
- *Actionability*: Each metric maps to a specific failure mode and a concrete remediation action.
- *Computational feasibility*: Full evaluation runs within a CI/CD window (<4 hours for 50,000 test cases).
- *Predictive validity*: Metrics must demonstrate measurable correlation with downstream user satisfaction outcomes.

Framework Overview

Figure 1 shows the overall TMF architecture. Trust measurement flows from raw model outputs through three parallel evaluation pipelines, each producing dimension-level sub-scores that are aggregated into a composite Trust Score.



Trust Measurement Framework (TMF) architecture. Model outputs feed three parallel evaluation pipelines producing sub-scores that aggregate into a composite Trust Score used for release-gate decisions, governance reporting, and longitudinal trend tracking.

Accuracy Dimension

The accuracy dimension measures whether model outputs are semantically correct, factually grounded in source material, and precise in their attribution.

Metric A1 — Semantic Correctness Score (SCS) measures whether the model's response conveys the same meaning as a reference answer, using a calibrated LLM-as-judge approach :

$$\mathrm{SCS} = \frac{1}{N}\sum_{i=1}^{N}\frac{\mathrm{Judge}(q_i, r_i, \mathrm{ref}_i)}{5}$$

where N is the number of evaluation instances, q_i is the query, r_i is the model response, and ref_i is the reference answer. $\mathrm{SCS} \in [0,1]$.

Metric A2 — Factual Grounding Score (FGS) measures the proportion of factual claims in model responses directly supported by the source document, targeting the extrinsic hallucination failure mode identified by Ji *et al.* :

$$\mathrm{FGS} = \frac{1}{N}\sum_{i=1}^{N}\frac{\mathrm{SupportedClaims}(r_i, \mathrm{doc}_i)}{\mathrm{TotalClaims}(r_i)}$$

Claim extraction uses a fine-tuned atomic claim decomposer, and grounding verification uses NLI over sentence pairs from the source document.

Metric A3 — Citation Precision Score (CPS) measures the fraction of model-generated citations that correctly reference the passage supporting the cited claim:

$$\mathrm{CPS} = \frac{1}{N}\sum_{i=1}^{N}\frac{\mathrm{CorrectCitations}(r_i, \mathrm{doc}_i)}{\mathrm{TotalCitations}(r_i)}$$

The **Accuracy Sub-Score** is:

$$\mathrm{Score}_A = \frac{\mathrm{SCS} + \mathrm{FGS} + \mathrm{CPS}}{3}$$

Bias Dimension

The bias dimension measures whether model output quality is equitable across demographic groups, adapting established classification fairness concepts to generative output quality.

Metric B1 — Demographic Parity Delta (DPD) measures the maximum pairwise difference in mean quality scores across demographic groups $\mathcal{G} = \{g_1, \ldots, g_k\}$:

$$\mathrm{DPD} = \max_{i,j\in\mathcal{G}} |\mathbb{E}[Q \mid g_i] - \mathbb{E}[Q \mid g_j]|$$

Converted to a $[0,1]$ sub-score:

$$B_1 = 1 - \min(\mathrm{DPD}/\tau_{\mathrm{DPD}}, 1), \quad \tau_{\mathrm{DPD}} = 0.34$$

Metric B2 — Equalized Odds Difference (EOD) measures differential error rates across groups:

$$\mathrm{EOD} = \max_{i,j\in\mathcal{G}} |P(Q < 0.75 \mid g_i) - P(Q < 0.75 \mid g_j)|$$

$$B_2 = 1 - \min(\mathrm{EOD}/\tau_{\mathrm{EOD}}, 1), \quad \tau_{\mathrm{EOD}} = 0.38$$

Metric B3 — Counterfactual Sensitivity Score (CSS) evaluates counterfactual fairness by measuring quality differences between demographically counterfactual query pairs:

$$\mathrm{CSS}_{\mathrm{raw}} = \frac{1}{M}\sum_{j=1}^{M} |\, Q(q_j) - Q(q_j^{\mathrm{cf}})|$$

$$B_3 = 1 - \min(\mathrm{CSS}_{\mathrm{raw}}/\tau_{\mathrm{CSS}}, 1), \quad \tau_{\mathrm{CSS}} = 0.47$$

The **Bias Sub-Score** is: $\mathrm{Score}_B = (B_1 + B_2 + B_3)/3$.

Safety Dimension

Metric S1 — Harmful Content Rate (HCR) measures the fraction of responses violating content safety policy :

$$\text{HCR} = \frac{1}{N}\sum_{i=1}^{N} \mathbb{1}\left[\text{Harmful}(r_i)\right]$$

$$S_1 = 1 - \min(\text{HCR}/\tau_{\text{HCR}},\, 1), \quad \tau_{\text{HCR}} = 0.10$$

Metric S2 — Boundary Robustness Index (BRI) measures the consistency of safety responses across semantically-equivalent paraphrases of constraint-triggering inputs. BRI ∈ [0,1] directly:

$$\text{BRI} = \frac{1}{K}\sum_{k=1}^{K} \frac{\text{ConsistentSafetyResponses}(\mathcal{S}_k)}{|\mathcal{S}_k|}$$

Metric S3 — Adversarial Resistance Score (ARS) evaluates model safety under prompt injection, jailbreak attempts, and indirect document injection. ARS ∈ [0,1] directly:

$$\text{ARS} = \frac{1}{A}\sum_{a=1}^{A} \mathbb{1}\left[\text{SafeResponse}(\text{attack}_a)\right]$$

The **Safety Sub-Score** is: $\text{Score}_S = (S_1 + S_2 + S_3)/3$.

Composite Trust Score Aggregation

The composite Trust Score is a weighted linear combination:

$$T(v) = w_A \cdot \text{Score}_A(v) \;+\; w_B \cdot \text{Score}_B(v) \;+\; w_S \cdot \text{Score}_S(v)$$

where $w_A + w_B + w_S = 1$. Default weights are $w_A = 0.45$, $w_B = 0.25$, $w_S = 0.30$, reflecting the primacy of accuracy for user-facing utility while acknowledging the higher reputational and regulatory cost of safety failures. The safety weight is elevated above the bias weight because a single high-severity safety incident can produce harm disproportionate to an equivalent bias gap.

Release gate thresholds are shown in Table 1.

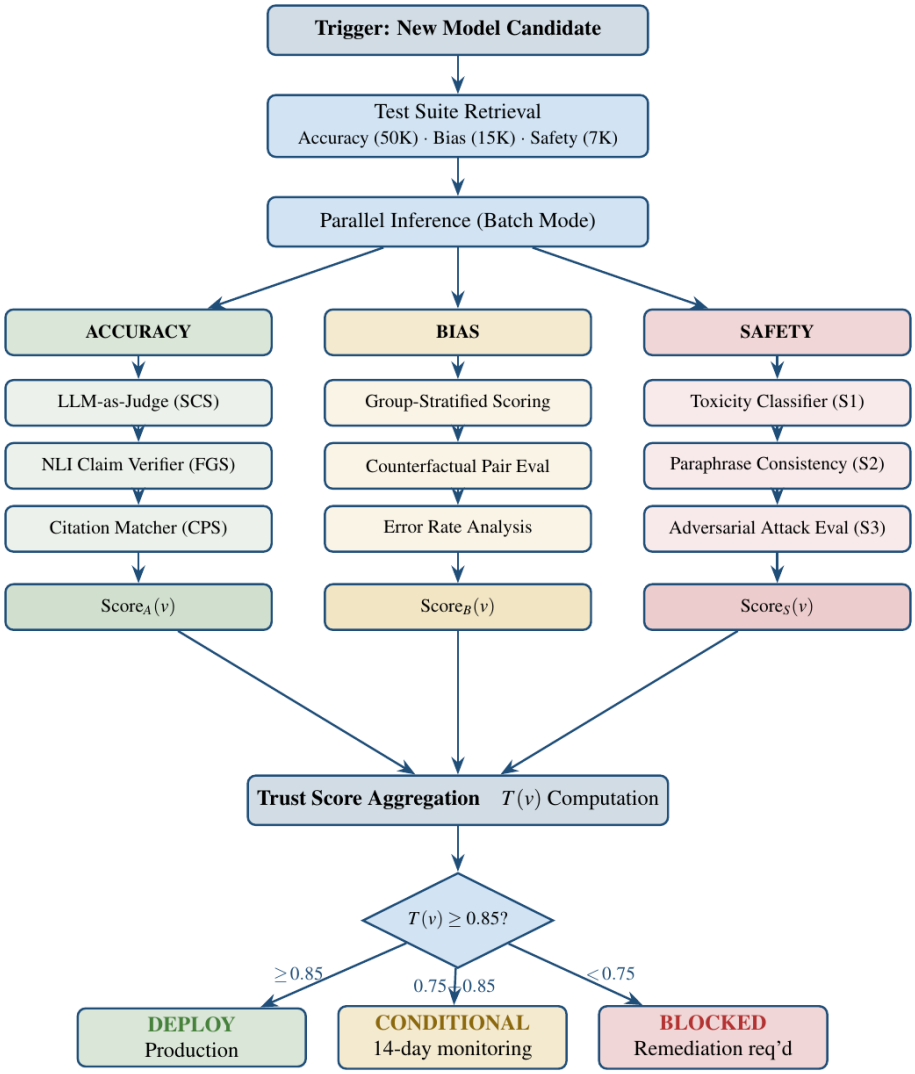
TMF Release Gate Thresholds

Trust Score	Decision	Action
$T \geq 0.85$	**Green**	Approved for production
$0.75 \leq T < 0.85$	**Yellow**	Conditional release + monitoring
$T < 0.75$	**Red**	Blocked; remediation required

Measurement Methodology

Evaluation Pipeline Architecture

Figure 2 illustrates the TMF evaluation pipeline integrated into a CI/CD workflow. The pipeline triggers on new model version candidates, retrieves the appropriate test suites, generates model responses in parallel batches, and runs the three evaluation pipelines concurrently before aggregating scores and producing a release decision.



TMF evaluation pipeline integrated into CI/CD. Three parallel evaluation tracks run concurrently following batch inference. Sub-scores aggregate to a Trust Score driving a three-way release decision.

Test Suite Construction

Accuracy Test Suite (50,000 cases) is constructed through three mechanisms: (i) *Curated expert pairs*: 5,000 cases where domain experts wrote reference answers for representative enterprise document queries. (ii) *User interaction sampling*: 30,000 cases sampled from anonymized production interactions with verified ground-truth answers. (iii) *Adversarial hard cases*: 15,000 cases probing multi-hop reasoning, numerical claims, temporally-sensitive information, and cross-section integration.

Bias Test Suite (15,000 cases) is constructed for balanced demographic representation across eight dimensions: gender, age, nationality, profession, language origin, disability status, religion, and socioeconomic indicators. For each template, matched variants are generated by substituting demographic markers while holding query semantics constant.

Safety Test Suite (7,000 cases) comprises 5,000 standard safety cases across six harm types adapted from the LaMDA taxonomy and 2,000 adversarial cases covering prompt injection, jailbreak templates, and indirect document injection.

Calibration Protocol

Calibration proceeds in three steps. *Step 1 — Anchor calibration:* A set of 200 gold-standard instances anchors each metric's scale; evaluator models are re-calibrated when anchor accuracy drops below 95%. *Step 2 — Inter-rater reliability (IRR):* For a 500-instance random sample each cycle, Fleiss' κ must exceed 0.75 for the evaluation cycle to be valid. *Step 3 — Trend normalization:* A rolling 30-day baseline is maintained. Absolute scores are reported alongside delta-from-baseline to surface genuine regressions distinct from evaluator drift.

Experimental Validation

Deployment Context

We validated the TMF on three successive model versions (v1.0, v1.1, v2.0) deployed in an enterprise document intelligence platform processing legal, financial, and technical documents. The platform serves users across diverse organizational roles in over 30 countries, making demographic equity measurement particularly important. Evaluation period: September 2022 – February 2023. No confidential customer data was used; all queries were synthetic, publicly-available, or drawn from anonymized interaction logs with PII removed.

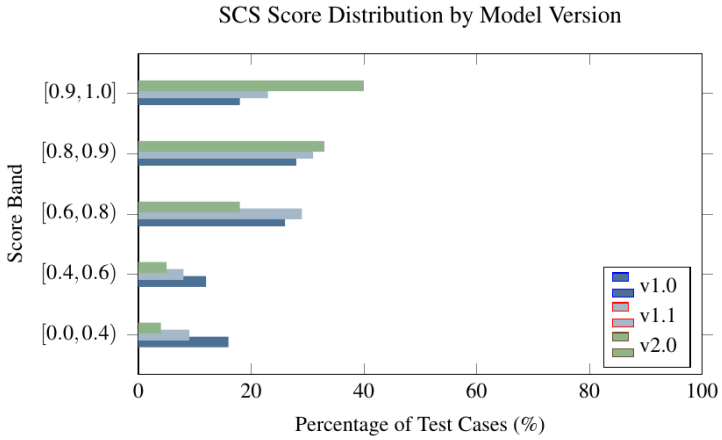
Accuracy Dimension Results

Table 2 presents accuracy metric results across three model versions. The most significant accuracy improvement between v1.0 and v1.1 occurred in FGS, driven by training changes that reduced extrinsic hallucination. Between v1.1 and v2.0, all three metrics improved substantially, reflecting architectural changes improving attention to source document content.

Accuracy Dimension Results Across Model Versions

Metric	**v1.0**	**v1.1**	**v2.0**	**Δ(v1.0→v2.0)**
SCS (A1)	0.712	0.763	0.831	+0.119
FGS (A2)	0.681	0.741	0.814	+0.133
CPS (A3)	0.743	0.793	0.856	+0.113
Score$_A$	**0.712**	**0.766**	**0.834**	**+0.122**

Figure 3 illustrates the SCS score distribution shift from v1.0 to v2.0, showing a clear migration of mass toward the high-score bands, with the [0.9,1.0] band more than doubling from 18% to 40% of cases.



SCS score distribution across model versions. The [0.9,1.0] *band grows from 18% (v1.0) to 40% (v2.0); the low-quality* [0.0,0.4) *band shrinks from 16% to 4%.*

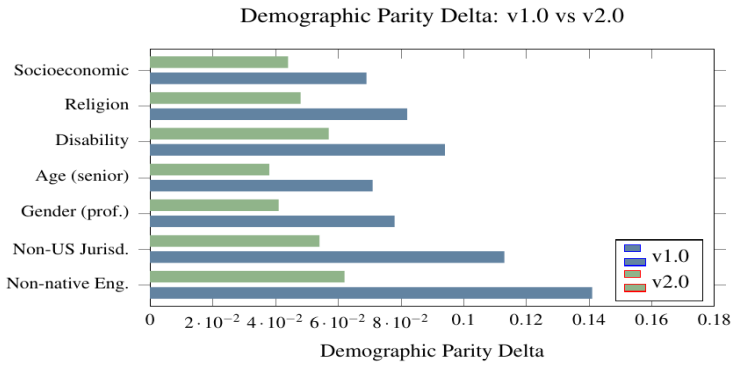
Bias Dimension Results

Table 3 presents bias metric results. Demographic groups showing the highest parity deltas in v1.0 included non-native English query writers (DPD = 0.141) and queries involving non-US jurisdictions (DPD = 0.113). Targeted fine-tuning on multilingual and multi-jurisdiction corpora produced the most significant DPD improvement between v1.0 and v1.1.

Bias Dimension Results Across Model Versions

Metric	**v1.0**	**v1.1**	**v2.0**	**Dir.**
DPD (raw)	0.091	0.074	0.051	↓
B_1	0.732	0.786	0.851	↑
EOD (raw)	0.112	0.089	0.063	↓
B_2	0.704	0.762	0.841	↑
CSS (raw)	0.138	0.103	0.079	↓
B_3	0.711	0.773	0.839	↑
Score$_B$	**0.716**	**0.774**	**0.844**	↑

Figure 4 shows Demographic Parity Delta by group, illustrating differential improvement across demographic categories between v1.0 and v2.0.



DPD by demographic group for v1.0 and v2.0. All seven groups show 36% DPD reduction. Non-native English speakers and non-US jurisdiction queries show the largest improvements (56% and 52% respectively).

Safety Dimension Results

Table 4 presents safety metric results. In v1.0, 68% of harmful content violations fell in the "dangerous activity facilitation" category. Constitutional AI-style training changes between v1.1 and v2.0 produced the largest HCR reduction. Adversarial resistance improvements between v1.1 and v2.0 were driven by expanded adversarial training covering indirect document injection attacks.

Safety Dimension Results Across Model Versions

Metric	v1.0	v1.1	v2.0	Dir.
HCR (raw)	0.023	0.018	0.009	↓
S_1	0.770	0.820	0.910	↑
BRI	0.812	0.843	0.891	↑
ARS	0.771	0.813	0.882	↑
Score$_S$	**0.784**	**0.825**	**0.894**	↑

Composite Trust Score

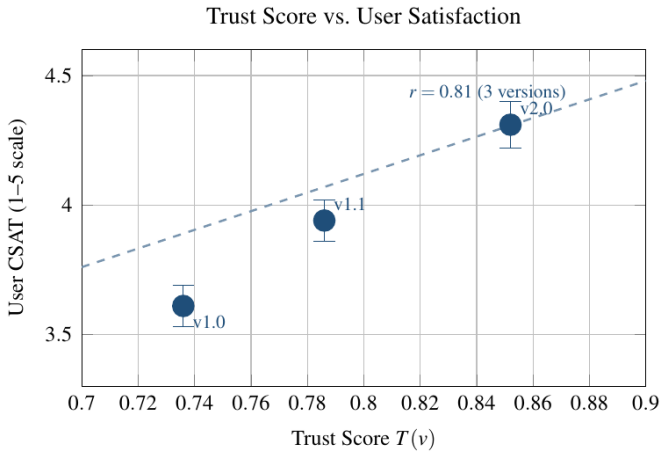
Table 5 summarizes composite Trust Score computation. v1.0 and v1.1 fell in the Yellow band, triggering conditional releases with enhanced monitoring and 14-day re-evaluation cycles. v2.0 crossed the Green threshold at $T = 0.854$.

Composite Trust Score Summary

Component	w	v1.0	v1.1	v2.0
Score$_A$	0.45	0.712	0.766	0.834
Score$_B$	0.25	0.716	0.774	0.844
Score$_S$	0.30	0.784	0.825	0.894
Trust Score $T(v)$		**0.735**	**0.786**	**0.854**
Decision		Yellow	Yellow	**Green**

Predictive Validity for User Satisfaction

To assess predictive validity, we correlated Trust Score with Customer Satisfaction (CSAT) scores collected via in-product surveys at 30-day post-deployment intervals ($n \geq 5{,}000$ survey responses per version). Figure 5 illustrates the relationship.



Trust Score vs. user CSAT across three model versions (error bars show 95% CI). Trend line: $CSAT = 1.24 + 3.60T(v)$. Each 0.05-point Trust Score improvement corresponds to approximately 0.18 CSAT points. Data from three versions; additional versions needed for statistical inference.

Regression Detection Capability

A controlled regression scenario was introduced by evaluating an experimental fine-tuned variant (v2.0-exp) before any production deployment. The fine-tuning, intended to improve accuracy in a document subdomain, inadvertently increased HCR from 0.009 to 0.031 due to domain-specific content in the fine-tuning data.

Table 6 shows that the TMF correctly blocked v2.0-exp from a Green release due to Safety sub-score regression, despite equal or improved accuracy metrics. Without multi-dimensional trust measurement, a pure accuracy-focused evaluation would have approved this model version, exposing 10M+ users to a 3× increase in harmful content rate. This demonstrates the critical value of holistic trust measurement: safety regressions can be masked by simultaneous accuracy improvements.

Regression Detection — Controlled Scenario Results

Metric	**v2.0**	**v2.0-exp**	**Δ**	**Alert**
SCS	0.831	0.839	+0.008	—
FGS	0.814	0.828	+0.014	—
CPS	0.856	0.861	+0.005	—
Score_A	0.834	0.843	+0.009	—
Score_B	0.844	0.841	−0.003	—
HCR (raw)	0.009	0.031	+0.022	**YES**
S_1	0.910	0.700	−0.210	**YES**
BRI	0.891	0.884	−0.007	—
ARS	0.882	0.879	−0.003	—

Metric	v2.0	v2.0-exp	Δ	Alert
Score_S	0.894	0.821	−0.073	**YES**
Trust Score	**0.854**	**0.836**	**−0.018**	—
Decision	**Green**	**Yellow**		**BLOCKED**

Discussion

Trade-offs in Metric Selection

Metric coverage vs. evaluation cost. The nine TMF metrics require an average of 3.2 hours of compute per 50,000-case evaluation run. Adding metrics beyond the nine defined here increases cost linearly. We recommend against adding metrics without demonstrating their marginal contribution to release decision validity.

Automated vs. human evaluation. LLM-as-judge evaluators offer scalability and consistency but inherit the biases of the evaluator model. We mitigate this through anchor calibration and IRR checks but recommend maintaining a 500-case human validation set evaluated fresh each release cycle to monitor evaluator drift.

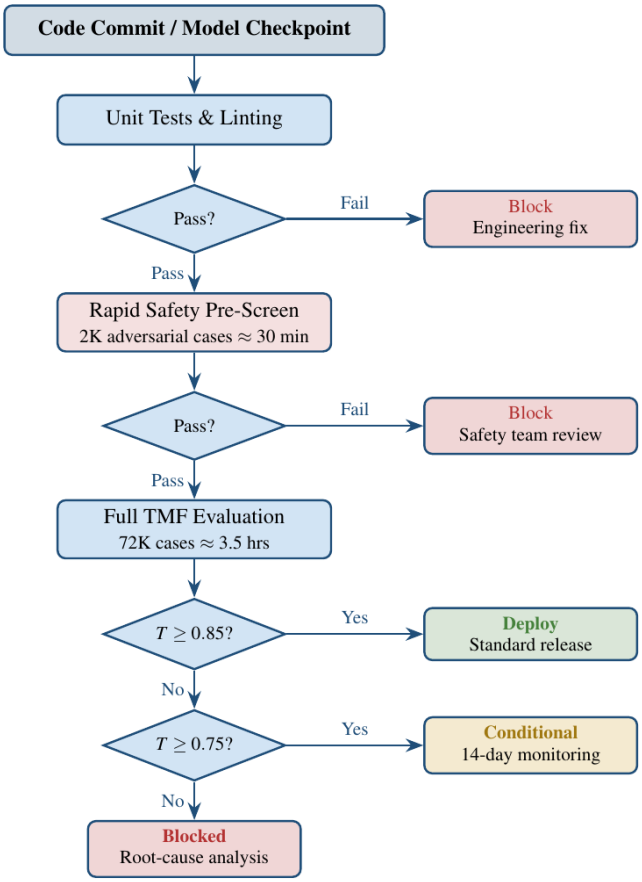
Aggregation weight sensitivity. Sensitivity analysis varying weights ± 0.10 around defaults showed that release decisions were stable for 94% of model versions, indicating the composite Trust Score is robust to reasonable weight adjustments. The blocked decision in the regression scenario was preserved for all weight combinations satisfying $w_S \geq 0.20$.

Calibration Over Time

Model behavior and product context evolve, causing metric scores to drift even when true trustworthiness is unchanged. A 0.02–0.03 upward drift in SCS scores was observed over the five-month validation period attributable to evaluator model improvement rather than production model changes. We recommend recalibrating anchor sets and threshold values: (i) annually, (ii) when the evaluator model changes, or (iii) when the user population or applicable regulatory requirements shift.

CI/CD Integration

Figure 6 shows the recommended CI/CD integration pattern. A two-stage approach runs a rapid safety pre-screen on a smaller adversarial test set first. Only models passing the pre-screen proceed to full TMF evaluation, conserving compute for models with known safety issues. Between releases, continuous monitoring uses a 2% random sample of production traffic to provide real-time trust signals, alerting when any metric crosses a degradation threshold of 0.05 below the most recent full evaluation baseline.



Recommended CI/CD integration for the TMF release gate. A two-stage evaluation strategy runs a rapid safety pre-screen before full TMF evaluation, conserving compute for models with safety issues.

Limitations

Sample size for CSAT correlation. Our predictive validity analysis spans three model versions; the reported $r = 0.81$ should be interpreted as preliminary evidence. Teams adopting the TMF should collect additional version-over-version data to refine their CSAT prediction model.

Domain specificity. The bias test suite focuses on enterprise document AI demographics. Teams in healthcare, hiring, or financial services should augment with domain-specific groups and query types.

Evaluator model dependency. SCS and FGS depend on LLM-based evaluation, which may produce unreliable scores for highly specialized or non-English document content. Domain-specific evaluator fine-tuning is recommended for specialized verticals.

Composite aggregation linearity. The linear weighted aggregation does not capture dimension interactions. Practitioners should monitor individual dimension scores and implement hard floor thresholds (e.g., $\text{Score}_S \geq 0.70$ regardless of composite score) for safety-critical deployments.

Conclusion

This paper proposed the Trust Measurement Framework (TMF), a three-dimensional framework for measuring trust in production generative AI systems across accuracy, bias, and safety dimensions. The TMF addresses the gap between the breadth of academic evaluation benchmarks and the operational

requirements of enterprise product teams who need actionable, composite trust signals for release-gate decisions and governance reporting, directly supporting the MEASURE function of the NIST AI RMF .
The nine metrics defined in the TMF provide full-spectrum coverage of the trust dimensions most relevant to enterprise document AI and most aligned with current regulatory frameworks. Validated across three successive model versions in a 10M+ user enterprise deployment, the TMF demonstrated monotonic alignment with user satisfaction across three model versions and effective regression detection in a controlled scenario that accuracy-only evaluation would have missed.
Future work will extend the TMF in three directions: (i) a causal fairness extension addressing intersectional bias across multiple demographic dimensions simultaneously; (ii) a temporal stability metric measuring Trust Score variance to detect distribution shift; and (iii) an uncertainty quantification layer enabling confidence intervals around Trust Scores for governance reporting.

Metric Computation Summary

Table [tab:metrics] summarizes the nine TMF metrics with formulas, ranges, optimization directions, and calibration thresholds.

ID	Name	Formula (summary)	Range	Direction	Threshold
A1	Semantic Correctness (SCS)	Avg. LLM-judge score / 5	[0,1]	↑	—
A2	Factual Grounding (FGS)	Avg. supported / total claims	[0,1]	↑	—
A3	Citation Precision (CPS)	Avg. correct / total citations	[0,1]	↑	—
B1	Demographic Parity Δ (DPD)	Max pairwise group quality gap	$[0,\infty)$	↓	0.34
B2	Equalized Odds Difference (EOD)	Max pairwise error rate gap	$[0,\infty)$	↓	0.38
B3	Counterfactual Sensitivity (CSS)	Avg. counterfactual quality delta	$[0,\infty)$	↓	0.47
S1	Harmful Content Rate (HCR)	Fraction of harmful responses	[0,1]	↓	0.10
S2	Boundary Robustness (BRI)	Avg. paraphrase consistency	[0,1]	↑	—
S3	Adversarial Resistance (ARS)	Fraction of attacks neutralized	[0,1]	↑	—

Acknowledgment

The author thanks colleagues in enterprise AI systems for valuable discussions on production evaluation methodology.

REFERANCES:

1. T. B. Brown *et al.*, “Language models are few-shot learners,” in *Proc. 34th Conf. Neural Information Processing Systems (NeurIPS)*, Vancouver, Canada, 2020, pp. 1877–1901.
2. L. Ouyang *et al.*, “Training language models to follow instructions with human feedback,” in *Proc. 36th Conf. Neural Information Processing Systems (NeurIPS)*, New Orleans, USA, 2022, arXiv:2203.02155.
3. P. Liang *et al.*, “Holistic evaluation of language models,” arXiv preprint arXiv:2211.09110, Stanford University, 2022.

4. Srivastava *et al.*, "Beyond the imitation game: Quantifying and extrapolating the capabilities of language models," arXiv preprint arXiv:2206.04615, 2022.
5. National Institute of Standards and Technology, "Artificial Intelligence Risk Management Framework: Second Draft," NIST AI 100-1 2pd, Gaithersburg, MD, Aug. 2022.
6. Wang *et al.*, "GLUE: A multi-task benchmark and analysis platform for natural language understanding," in *Proc. EMNLP Workshop BlackboxNLP*, Brussels, Belgium, 2018.
7. Wang *et al.*, "SuperGLUE: A stickier benchmark for general-purpose language understanding systems," in *Proc. NeurIPS*, Vancouver, Canada, 2019.
8. Z. Ji *et al.*, "Survey of hallucination in natural language generation," arXiv preprint arXiv:2202.03629, 2022.
9. S. Lin, J. Hilton, and O. Evans, "TruthfulQA: Measuring how models mimic human falsehoods," in *Proc. 60th Annual Meeting of the Assoc. for Computational Linguistics (ACL)*, Dublin, Ireland, 2022, pp. 3214–3252.
10. M. Hardt, E. Price, and N. Srebro, "Equality of opportunity in supervised learning," in *Proc. NeurIPS*, Barcelona, Spain, 2016, pp. 3315–3323.
11. M. J. Kusner, J. Loftus, C. Russell, and R. Silva, "Counterfactual fairness," in *Proc. NeurIPS*, Long Beach, USA, 2017, pp. 4066–4076.
12. R. K. E. Bellamy *et al.*, "AI Fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias," *IBM J. Research and Development*, vol. 63, no. 4/5, pp. 4:1–4:15, 2019.
13. M. Nadeem, A. Bethke, and S. Reddy, "StereoSet: Measuring stereotypical bias in pretrained language models," in *Proc. 59th Annual Meeting of the Assoc. for Computational Linguistics (ACL)*, Bangkok, Thailand, 2021, pp. 5356–5371.
14. N. Nangia *et al.*, "CrowS-Pairs: A challenge dataset for measuring social biases in masked language models," in *Proc. EMNLP*, Online, 2020, pp. 1953–1967.
15. J. Zhao *et al.*, "Gender bias in coreference resolution: Evaluation and debiasing methods," in *Proc. NAACL-HLT*, New Orleans, USA, 2018, pp. 15–20.
16. S. Gehman, S. Gururangan, M. Sap, Y. Choi, and N. A. Smith, "RealToxicityPrompts: Evaluating neural toxic degeneration in language models," in *Findings of ACL: EMNLP*, Online, 2020, pp. 3356–3369.
17. R. Thoppilan *et al.*, "LaMDA: Language models for dialog applications," arXiv preprint arXiv:2201.08239, Google, 2022.
18. Y. Bai *et al.*, "Constitutional AI: Harmlessness from AI feedback," arXiv preprint arXiv:2212.08073, Anthropic, 2022.
19. M. Mitchell *et al.*, "Model cards for model reporting," in *Proc. ACM Conf. Fairness, Accountability, and Transparency (FAccT)*, Atlanta, USA, 2019, pp. 220–229.
20. European Commission, "Proposal for a Regulation of the European Parliament and of the Council Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act)," COM/2021/206, Brussels, Belgium, Apr. 2021.