Amended Proposal for AI Risk Scorecard Assessment

Ethan Barrilleaux, Jada Battle, Lambert Li, Aidan Whitlock, and Makeenie Robinson
Wake Forest University

MSBA x Mastercard Practicum Project

User Interaction / Satisfaction Metrics

1.1 Hallucination Rate Reduction

Recommendation: Implement Retrieval-Augmented Generation (RAG) and confidence scoring to minimize hallucinated AI responses.

(Prototype question for Qualtrics)

> Hallucination Rate Control: How frequently does your AI model generate incorrect or misleading outputs? (Scale: Never - Very Often)
>
> - Please provide the latest measured Hallucination Rate. See a sample from the appendix.

1.2 User Override Rate Optimization

Recommendation: Use SHAP-based explainability tools to improve AI recommendation acceptance. Assess AI-response quality from User Override Rate (if any).

> How do you currently assess user confidence in AI-driven decisions? (Multiple Choice: Post-interaction surveys, Override monitoring, Qualitative feedback, No structured assessment, Other)
>
> User Override Rate Assessment: How often do human reviewers override AI-generated recommendations? (Scale: Never - Very Often)
>
> - Please provide the User Override Rate percentage. See a sample from appendix.

1.3 AI-Driven Value Attribution Score Tracking

Recommendation: Measure the impact of AI-driven decisions on business outcomes such as cost savings and revenue growth.
Proposal: Deploy A/B testing frameworks to track and compare AI-driven vs. human-driven results.

> Business Impact Validation: Can you provide quantifiable evidence of AI-driven improvements in business performance? (Yes/No)
>
> - If Yes, please provide supporting metrics (e.g., revenue impact, cost reduction, user engagement improvement).

1.4 Decision Confidence Score (DCS) Implementation

Recommendation: Implement a Decision Confidence Score (DCS) to measure user trust and confidence in AI-generated recommendations. Integrate DCS tracking into Mastercard's AI governance pipeline through post-interaction surveys and override monitoring.

User Trust & Risk Assessment:

- How confident are you in the AI-generated recommendations / results? (Scale: Not at all – Completely confident)
- Have you overridden an AI recommendation due to a lack of confidence? (Yes/No)

1.5 Friction Index (FI) Integration

Recommendation: Implement the Friction Index (FI) to quantify inefficiencies and pain points in AI interactions, ensuring smoother and more efficient user experiences.

Proposal: Track user journey data, including task completion times, abandonment rates, and error tracking to identify areas of friction in AI-driven systems.

User Interaction & Friction Assessment:

- How frequently do you encounter difficulties or delays when interacting with AI-powered systems? (Scale: Never – Very Often)

2. Model Metrics

2.1 Perplexity & Long Perplexity (PPL) Management

Recommendation: Monitor perplexity scores to detect model inconsistencies and ensure stable, coherent AI-generated text.

Model Perplexity Evaluation: How well does your AI model maintain coherence in long-form text generation? (Scale: Poor - Excellent)

- Please provide the most recent Perplexity Score and Long Perplexity Score.

2.2 Prediction Entropy Reduction

Recommendation: Lower AI uncertainty by refining probability thresholds and incorporating entropy-based filtering.

Prediction Confidence Monitoring: How does your AI system handle high-entropy (uncertain) predictions? (Multiple Choice: Automatic Review, Human Override, No Monitoring, Other)

- Please provide the latest Prediction Entropy Score.

2.3 Kullback-Leibler (KL) & Jensen-Shannon (JS) Divergence Stability Tracking

Recommendation: Monitor KL and JS divergence to detect distribution shifts in AI-generated responses.
Proposal: Implement automated drift detection for high-risk AI applications to ensure stability.

Data Distribution Monitoring: How do you track changes in model input distributions over time? (Multiple Choice: KL Divergence, JS Divergence, Manual Checks, No Tracking, Other)

- Please provide the most recent KL Divergence and JS Divergence Scores.

2.4 Perplexity Drift Rate Monitoring

Recommendation: Track perplexity drift to maintain language model stability in customer support and financial reporting applications.
Proposal: Define acceptable perplexity thresholds and trigger retraining when deviations occur.

Do you have a definite threshold in place that triggers model retraining at a certain Perplexity Score? (Yes/No)

- If yes, what is that defined threshold? (Numeric Input)

AI Model Retraining Strategy: How frequently do you retrain AI models to mitigate performance drift? (Scale: Never - Monthly)

- Please provide the last retraining date and any observed model drift scores.

2.5 Response Consistency Score (RCS) Monitoring

Recommendation: Monitor the Response Consistency Score (RCS) to ensure that AI-generated responses remain stable, contextually appropriate, and reliable over time.

Proposal: Track AI-generated responses for consistency and context alignment by analyzing response audits and user feedback.

Consistency & Accuracy Assessment:

> How consistent are AI-generated responses across multiple interactions on the same or similar topics? (Scale: Very Inconsistent – Very Consistent)

Appendix: