Generative AI ROI and Productivity Assessment

Presentation for Mastercard AI Governance Unit

04/28/2025

Ethan Barrilleaux, Jada Battle, Lambert Li, Makeenie Robinson, Aidan Whitlock

Wake Forest MSBA Practicum Team

MEET OUR TEAM

Makeenie Robinson

Team Lead

Jada Battle

Client Liaison

Aidan Whitlock

Advisor Liaison

Ethan Barrilleaux

Design Lead

Lambert Li

Research and Knowledge Officer

WAKE FOREST BUSINESS

AGENDA

01

Context and Insights

- Gaps in current scorecard for GenAI ROI
- Industry flags: ROI missteps & risk blind spots
- Need for business-aligned metrics

02

Our Solution

- User trust + model performance metrics
- Logic-embedded questions, audience clarity
- Anchoring fairness, efficacy, transparency

03

Final Deliverables

- Actionable, logic-driven question set
- Excel tracker + Qualtrics survey demo
- Clear guidance for internal + vendor rollout

WAKE FOREST BUSINESS

WFU MSBA x Mastercard | Generative AI ROI & Risk Scorecard Project

A Smarter Scorecard to Drive GenAI ROI, Mitigate Risk, and Improve Outcomes

1. OVERVIEW

CHALLENGE

Existing scorecard lacks GenAI ROI granularity

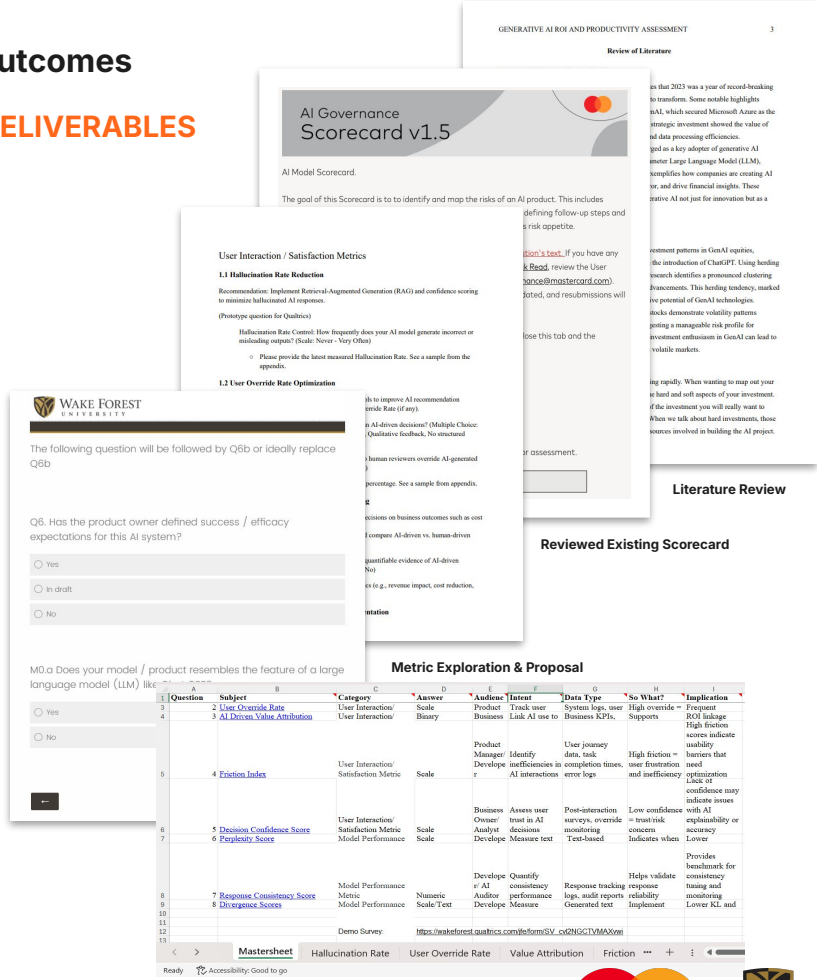
SOLUTION

Developed targeted, business-aligned metrics with survey logic

IMPACT

Supports better decision-making, risk reduction, and value tracking

2. DELIVERABLES



Literature Review

Reviewed Existing Scorecard

Metric Exploration & Proposal

Excel Masterlist

3. CONTEXT

Narrow Scope
Existing scorecards prioritize technical risk over business value

GenAI Blind Spots
Key GenAI risks (hallucination, drift, bias) remain under-assessed

Principle-Driven Need
Mastercard's principles call for measurable fairness, efficacy, transparency

Usability Gap
Business owners need clear, actionable questions—not just technical checks

WAKE FOREST BUSINESS

From Kickoff to Delivery: Our End-to-End Scorecard Development Timeline (Oct. 2024–Apr. 2025)

Oct. - Nov. 2024

Literature Review

High-level research from industry-leader sources

Jan. 2025

Existing Scorecard Evaluation

Form ideas on additional metrics and our begin creating our own recommendations

Jan. - Feb. 2025

Metric Exploration

Measurability, data sources, integration, and challenges/considerations

Feb. 2025

Proposal and Qualtrics Survey Draft

Wrote amended proposal with actual questions we could add to scorecard + Qualtrics survey

Mar. 2025

RAG Workshop and Feedback Integration

Refined questions and understood more how Mastercard could adopt them

Apr. 2025

Final Presentation

Right now!

WAKE FOREST BUSINESS

An overview from the literature on Generative AI's business value and productivity impact.

1

Framing the Challenge

To ground our scorecard in evidence, we reviewed literature on GenAI's impact on productivity and ROI.

2

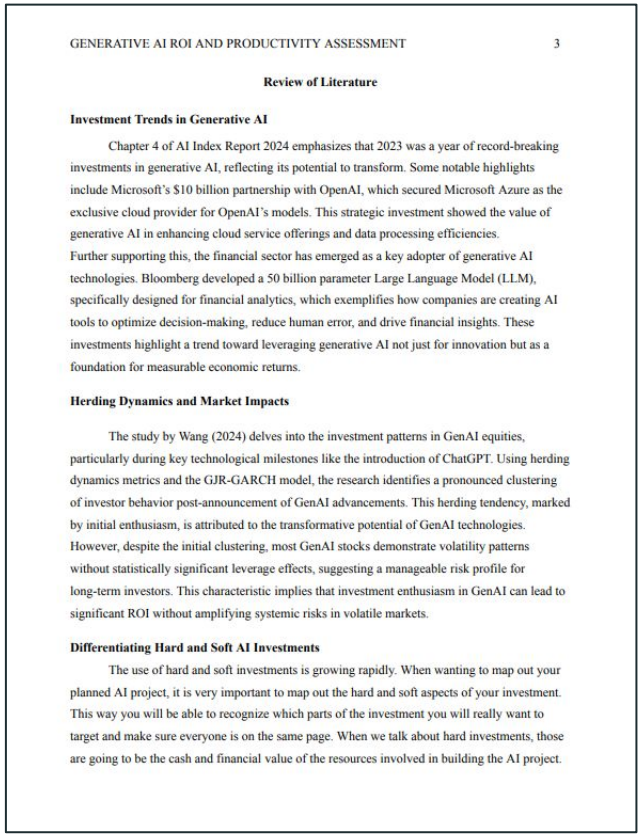
Key Findings

Research shows GenAI can boost workflows and decision-making, but most organizations lack clear ways to measure its value.

3

How It Shaped Our Work

This shaped how we built our scorecard revealing gaps, refining questions, and aligning metrics to Mastercard's GenAI goals.



GENERATIVE AI ROI AND PRODUCTIVITY ASSESSMENT 3

Review of Literature

Investment Trends in Generative AI

Chapter 4 of AI Index Report 2024 emphasizes that 2023 was a year of record-breaking investments in generative AI, reflecting its potential to transform. Some notable highlights include Microsoft's $10 billion partnership with OpenAI, which secured Microsoft Azure as the exclusive cloud provider for OpenAI's models. This strategic investment showed the value of generative AI in enhancing cloud service offerings and data processing efficiencies.
Further supporting this, the financial sector has emerged as a key adopter of generative AI technologies. Bloomberg developed a 50 billion parameter Large Language Model (LLM), specifically designed for financial analytics, which exemplifies how companies are creating AI tools to optimize decision-making, reduce human error, and drive financial insights. These investments highlight a trend toward leveraging generative AI not just for innovation but as a foundation for measurable economic returns.

Herding Dynamics and Market Impacts

The study by Wang (2024) delves into the investment patterns in GenAI equities, particularly during key technological milestones like the introduction of ChatGPT. Using herding dynamics metrics and the GJR-GARCH model, the research identifies a pronounced clustering of investor behavior post-announcement of GenAI advancements. This herding tendency, marked by initial enthusiasm, is attributed to the transformative potential of GenAI technologies. However, despite the initial clustering, most GenAI stocks demonstrate volatility patterns without statistically significant leverage effects, suggesting a manageable risk profile for long-term investors. This characteristic implies that investment enthusiasm in GenAI can lead to significant ROI without amplifying systemic risks in volatile markets.

Differentiating Hard and Soft AI Investments

The use of hard and soft investments is growing rapidly. When wanting to map out your planned AI project, it is very important to map out the hard and soft aspects of your investment. This way you will be able to recognize which parts of the investment you will really want to target and make sure everyone is on the same page. When we talk about hard investments, those are going to be the cash and financial value of the resources involved in building the AI project.

WAKE FOREST BUSINESS

Analyzing the existing scorecard to identify gaps and inform targeted improvements.

✓ We identified gaps in Mastercard's risk scorecard around GenAI. Our solution: tailored, clear, and targeted questions that better align with GenAI risks and business needs.

Existing Approach	Identified Gaps	Our Approach
Post-build evaluation focus	Misses early stage risk planning and alignment	Integrated Pre/Post build tags to clarify question timing and relevance
Broad, AI generic questions	Fails to capture the unique challenges of GenAI	Developed GenAI specific metrics
Technical and complex phrasing	Difficult for business owners to interpret and act upon	Applied audience mapping and simplified language for clarity

Understand how metric categories drive design philosophy.

Why Separate Them?

- ✓ **A GenAI system may perform very well technically, but have poor ROI if employees do not use, understand, or trust the model.**

Recommendation: Implement a Decision Confidence Score (DCS) to me

confidence in AI-generated recommendations. Integrate DCS tracking in

governance pipeline through post-interaction surveys and override monit

User Trust & Risk Assessment:

- How confident are you in the AI-generated recommendati Not at all – Completely confident)
- Have you overridden an AI recommendation due to a lack (Yes/No)

1.5 Friction Index (FI) Integration

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

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

User Interaction & Friction Assessment:

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

2. Model Metrics

2.1 Perplexity & Long Perplexity (PPL) Management

Recommendation: Monitor perplexity scores to detect model inconsisten 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

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

WAKE FOREST BUSINESS

Incorporate metrics to better quantitative measurements.

1 What Did We Do?

We designed a scorecard and survey by identifying key AI metrics, mapping them to specific questions, and organizing everything into a master Excel document for streamlined analysis.

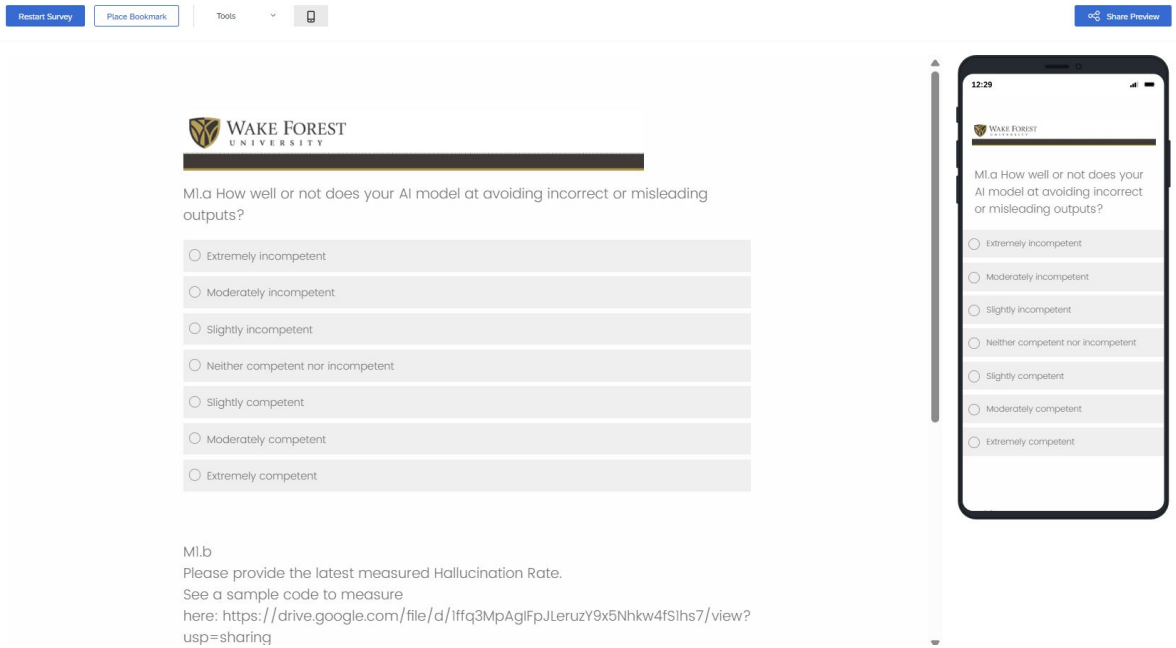
Question Identifier	Subject	Category	Answer Format	Audience	Intent	Data Type Needed	So What?	Implication
1	Hallucination Rate	User Interaction/ Satisfaction Metric	Scale	Business Owner/ Developer	To assess the factual reliability of GenAI outputs	Model outputs, benchmark facts, user feedback	High hallucination = trust/risk concern	High hallucination rate implies misinformation, reputational risk, and low adoption
2	User Override Rate	User Interaction/ Satisfaction Metric	Scale	Product Manager / Analysts	Track user trust & alignment between AI suggestions and human judgment	System logs, user flags, override actions	High override = low trust	Frequent overrides may signal explainability or accuracy issues
3	AI Driven Value Attribution	User Interaction/ Satisfaction Metric	Binary	Business Owner/ Developer	Link AI use to measurable ROI	Business KPIs, sales data, operational cost data	Supports investment decisions	ROI linkage enables prioritization of AI projects
4	Friction Index	User Interaction/ Satisfaction Metric	Scale	Product Manager / Developer	Identify inefficiencies in AI interactions	User journey data, task completion times, error logs	High friction = user frustration and inefficiency	High friction scores indicate usability barriers that need optimization
5	Decision Confidence Score	User Interaction/ Satisfaction Metric	Scale	Business Owner/ Analyst	Assess user trust in AI decisions	Post-interaction surveys, override monitoring	Low confidence = trust/risk concern	Lack of confidence may indicate issues with AI explainability or accuracy
6	Perplexity Score	Model Performance Metric	Scale	Developer/AI Auditor	Measure text coherence among model iterations	Text-based model output, benchmark (industry standard)	Indicates when LLM may start to deviate from coherent text generation	Lower perpelxity score = more coherent and stable AI generated text
7	Response Consistency Score	Model Performance Metric	Numeric	Developer/ AI Auditor	Quantify consistency performance	Response tracking logs, audit reports	Helps validate response reliability	Provides benchmark for consistency tuning and monitoring
8	Divergence Scores	Model Performance Metric	Scale/Text	Developer/AI Auditor	Measure content divergence across model iterations	Generated text from LLM and industry standard reference text	Implement automated drift detection for model output	Lower KL and JS Divergence scores = less drift

2 From Metrics to Questions

Question Identifier	Subject	Category	Question Body	Potential Answers	Answer Format	Audience	Pre/ Post?	Required to answer?	Condition	Preceding question #	Next question #	Intent	Data Type Needed	So What?	Implication
M1.a	Hallucination Rate	User Interaction/ Satisfaction Metric	How well does your AI model at avoiding incorrect or misleading outputs?	Extremely incompetent to extremely competent (7 point scale)	Multiple Choice	Business Owner/ Developer	Post	Yes	If yes in Q33	Q6b.What potential metrics do you intend to use to track success?		To assess the factual reliability of GenAI outputs	Model outputs, benchmark facts, user feedback	High hallucination = trust/risk concern	High hallucination rate implies misinformation, reputational risk, and low adoption

WAKE FOREST BUSINESS

Sample Implementation:
Bring The Scorecard To Life In Qualtrics



https://wakeforest.qualtrics.com/jfe/form/SV_cvI2NGCTVMAXvwi

WAKE FOREST BUSINESS

Visualize how questions will be added to Mastercard framework.

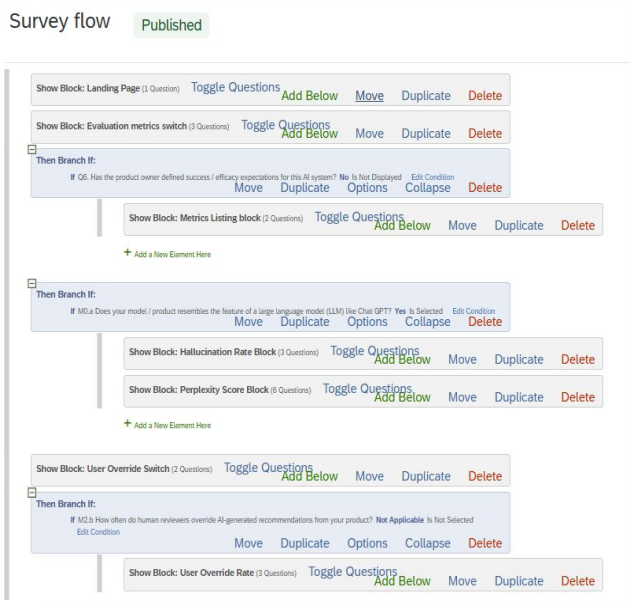
1 Clear Placement w.r.t Scorecard

Identify existing scorecard area that have attempted to capture quantitative metrics and supplement with additional measurements in:

- User-centric / end product metrics
- Model performance metrics

Provide added clarity on how proposed metrics and merge seamlessly with existing system.

2 Conditional Logic



Incorporate conditional logic to better target applicable survey users to dedicated questions:

- Ensure accurate metrics being highlighted and areas of interested focused.
- Raise awareness on business metrics in early product dev.
- Prevent irrelevant and inaccurate information gathering.

WAKE FOREST BUSINESS

Visualize how questions will be added to Mastercard framework.

3 Implementation Sample

M3.d If you have supporting data, please provide the friction index by referencing the following
sample: https://drive.google.com/file/d/10QZGk9kHshP_7WO0FyB4X4vdRlLtAH93/view?usp=sharing

```
class FrictionIndexMonitor:
    def __init__(self):
        self.error_counts = []  # Tracks AI recommendation errors
        self.resolution_times = []  # Time taken to resolve AI-related issues
        self.drop_off_rates = []  # Tracks user drop-off rates due to AI inefficiencies

    def record_interaction(self, errors, resolution_time, drop_off_rate):
        """Records an AI interaction event for friction index calculation."""
        self.error_counts.append(errors)
        self.resolution_times.append(resolution_time)
        self.drop_off_rates.append(drop_off_rate)

    def calculate_friction_index(self):
        """Computes the Friction Index (FI) as a weighted score."""
        if not self.error_counts:
            return None

        avg_errors = sum(self.error_counts) / len(self.error_counts)
        avg_resolution_time = sum(self.resolution_times) / len(self.resolution_times)
        avg_drop_off = sum(self.drop_off_rates) / len(self.drop_off_rates)

        # Assigning weights to factors (customizable based on impact analysis)
        fi_score = (0.4 * avg_errors) + (0.3 * avg_resolution_time) + (0.3 * avg_drop_off)
        return fi_score
```

Create sample implementation code to run business-centered test to help complete the scorecard.

4 Interpretable Metrics

So What?	Implication
High hallucination = trust/risk concern	High hallucination rate implies misinformation, reputational risk, and low adoption
High override = low trust	Frequent overrides may signal explainability or accuracy issues
Supports investment decisions	ROI linkage enables prioritization of AI projects
High friction = user frustration and inefficiency	High friction scores indicate usability barriers that need optimization

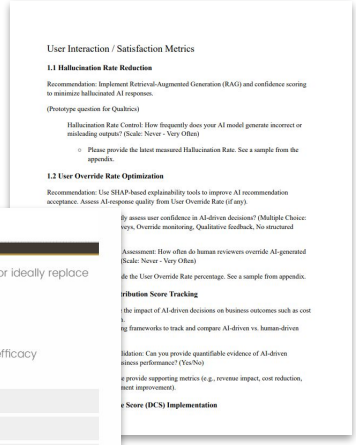
Each question is followed by a note of metrics result. More explorations can be done for setting up successful benchmarks from past projects.

Currently, no industry standard on particular scores.

WAKE FOREST BUSINESS

Final Outputs Designed to Support Deployment and Decision-Making

- ✓ **Master Excel Sheet of 20+ Tagged Questions**
- ✓ **Built-in Conditional Logic**
- ✓ **Demo Survey in Qualtrics**
- ✓ **Documents With Metrics, Use Cases, and Definitions**



User Interaction / Satisfaction Metrics

1.1 Hallucination Rate Reduction

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

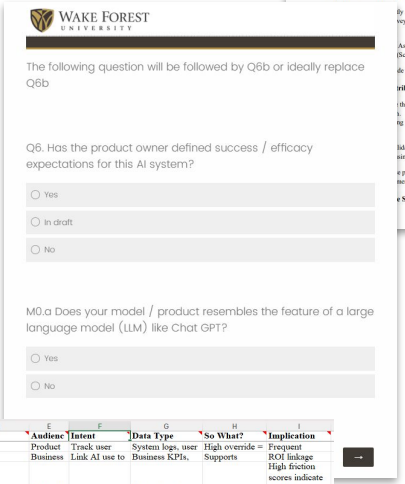
(Prototype question for Qualtrics)

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).



WAKE FOREST UNIVERSITY

The following question will be followed by Q6b or ideally replace Q6b

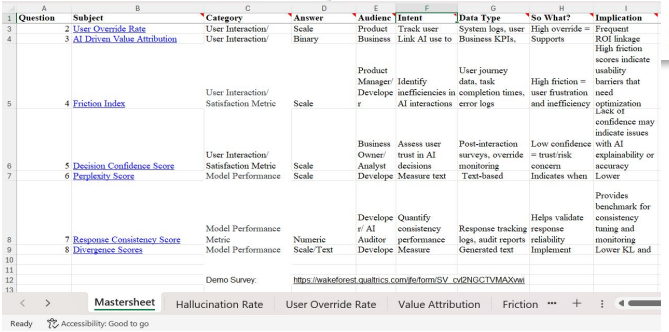
Q6. Has the product owner defined success / efficacy expectations for this AI system?

- Yes
- In draft
- No

M0.a Does your model / product resembles the feature of a large language model (LLM) like Chat GPT?

- Yes
- No

→



	A	B	C	D	E	F	G	H	I
1	Question	Subject	Category	Answer	Audienc	Intent	Data Type	So What?	Implication
3	2	User Override Rate	User Interaction/	Scale	Product	Track user	System logs, user	High override =	Frequent
4	3	AI Driven Value Attribution	User Interaction/	Binary	Business	Link AI use to	Business KPIs,	Supports	ROI linkage
5	4	Friction Index	User Interaction/ Satisfaction Metric	Scale	Product Manager/ Developer	Identify inefficiencies in AI interactions	User journey data, task completion times, error logs	High friction = user frustration and inefficiency	High friction scores indicate usability barriers that need optimization
6	5	Decision Confidence Score	User Interaction/ Satisfaction Metric	Scale	Business Owner/ Analyst	Assess user trust in AI decisions	Post-interaction surveys, override monitoring	Low confidence = trust/risk concern	Lack of confidence may indicate issues with AI explainability or accuracy
7	6	Perplexity Score	Model Performance	Scale	Develope	Measure text	Text-based	Indicates when	Lower
8	7	Response Consistency Score	Model Performance Metric	Numeric	Developer/ AI Auditor	Quantify consistency performance	Response tracking logs, audit reports	Helps validate response reliability	Provides benchmark for consistency tuning and monitoring
9	8	Divergence Scores	Model Performance	Scale/Text	Develope	Measure	Generated text	Implement	Lower KL and
10									
11									
12			Demo Survey:	https://wakeforest.qualtrics.com/jfe/form/SV_cvl2NGCTVMAXvwi					
13									

Mastersheet | Hallucination Rate | User Override Rate | Value Attribution | Friction

Ready | Accessibility: Good to go

WAKE FOREST BUSINESS

Appendix

1. **Literature Review**: Literature Review
2. **Metrics Written Report**: Metrics Written Report
3. **Amended Proposal**: Amended Proposal
4. **Question Design (Qualtrics Survey)**: Qualtrics Survey
5. **Final Excel Spreadsheet**: Final Deliverable

WAKE FOREST BUSINESS