USD(R&E)

Hazard Analysis for RAG-LLM Systems

September 2025

Elena Charnetzki
Carol Pomales
MITRE Support
Developmental Test, Evaluation, and Assessments

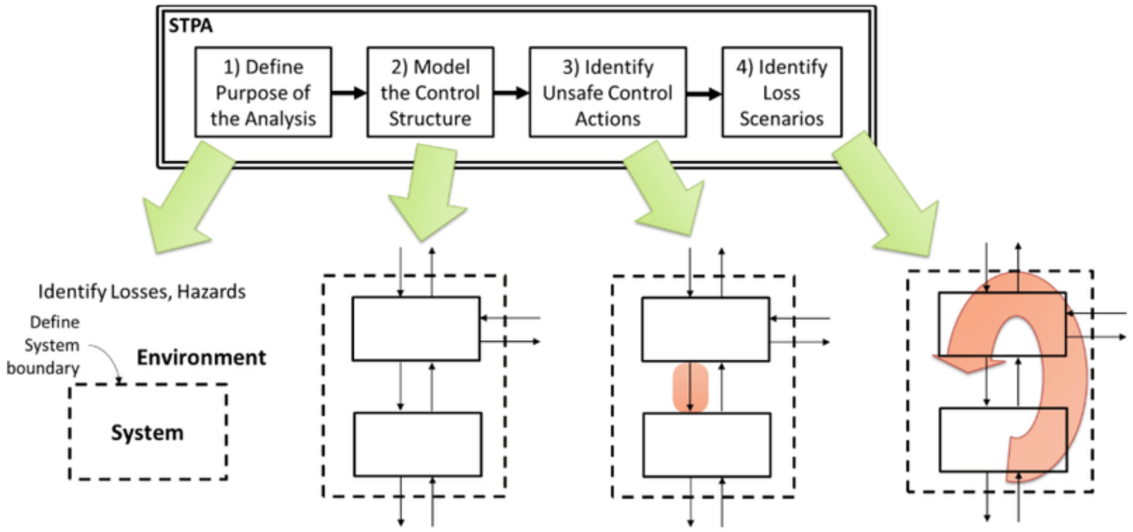
Controlled by: D(DTE&A), OUSD(R&E)
CUI Category: n/a
Distribution: Distribution A – Unlimited; DOPSAR Case 25-T-3151
POC: Mr. Orlando Flores, 571.3724145

Overview

- MITRE is working in support of the OUSD(R&E), DTE&A to improve and enable T&E strategies for Generative AI, which is seeing increasing adoption in capabilities across DoD.
- We will discuss this method framed in the context of a fictional intelligence analysis, that provides hypothesis for consideration.
- We will show how one can use an MBSE-based approach to implement a hazard analysis methodology, Systems Theoretic Process Analysis (STPA).
- Metrics used to evaluate the AIES output against expert-generated ground truth answers or User SMEs with Trust Metrics can be effective.
- The goal is to ensure that robust, relevant, and adaptable processes are establish to enable the challenges with this form of AI technology.

STPA and Hazard Scenarios

- STPA (Systems Theoretic Process Analysis) is a method that has historically been used to identify hazards in complex systems (1).
- We applied STPA and integrated it with MBSE to model hazards of generative AI drawn from key literature (2)(3)(4)
- This allowed us to identify six archetypal critical hazard scenarios that represent the most common worst-case scenarios.
- These critical hazard scenarios can be used as a reference model to enable customized hazard analysis.



Critical Hazard Scenarios
Malicious user succeeds in generating malicious content
User engages in unauthorized use
Capability generates unacceptable content in response to a benign prompt
Unacceptable quality output goes undetected
User is unable to correct unacceptable quality output
User is over- or under-reliant on system

1 Schulker, David. *Using System Theoretic Process Analysis to Advance Safety in LLM-enabled Software Systems*, 2024.
2 NIST AI 600-1, *Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile*, 2024.
3 MITRE Risk Discovery Protocol for AI Assurance
4 Li et al, *A Closer Look at the Existing Risks of Generative AI: Mapping the Who, What, and How of Real World Incidents*, 2025.

Approach to STPA for Enabling RAG-LLMs

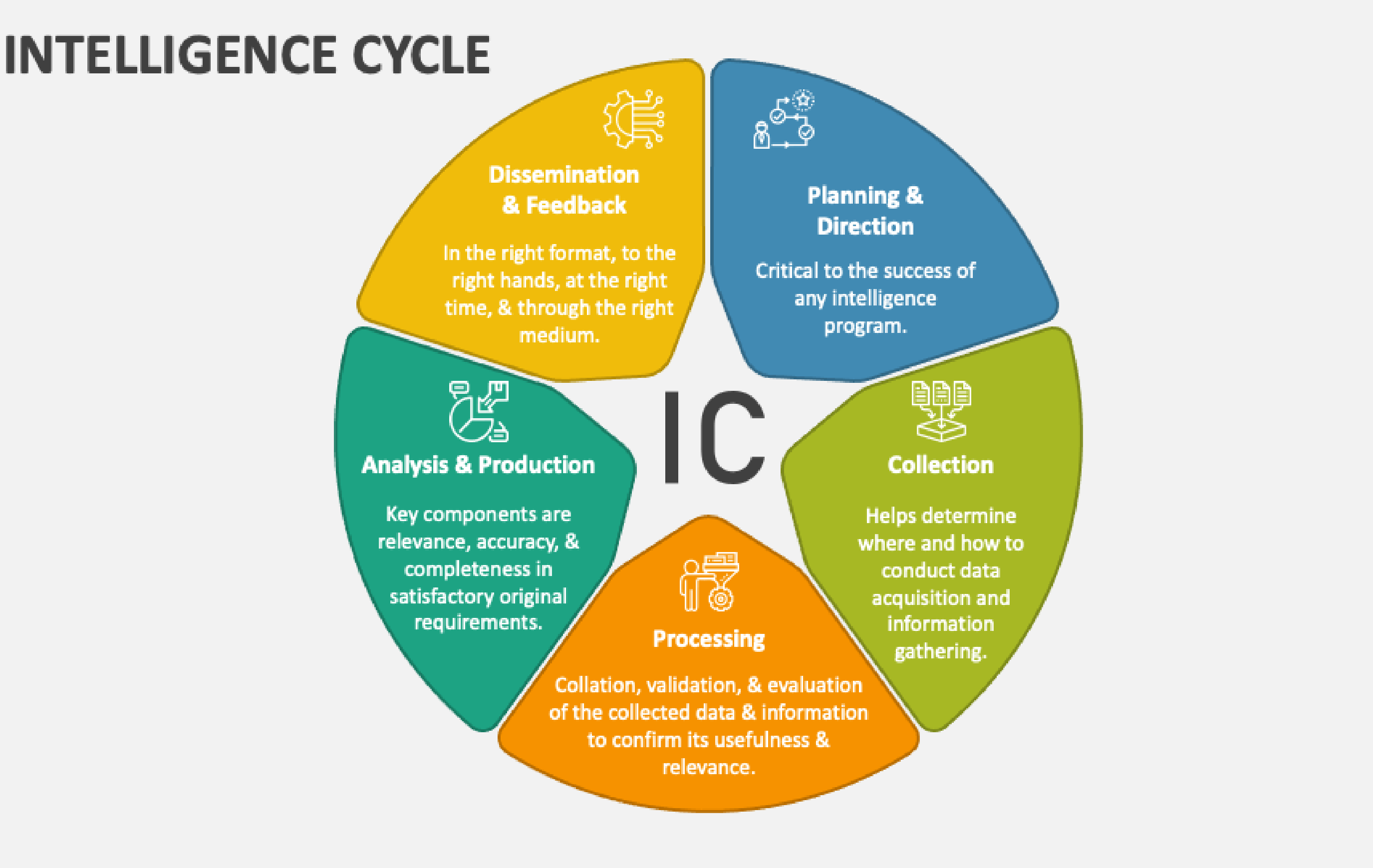
Challenge Encountered:	Solution Developed:
Difficulty. Generating a sufficiently complete set of potential hazards from a "blank page" is very difficult and time consuming, even for AI and domain subject matter experts	**Generalized Models.** Developed a reference activity model for RAG-LLM systems to use MBSE to quickly place the hazard in the right mission context and explore hazard propagation scenarios systematically, making the hazard analysis more complete while saving time
Efficiency. Engagements with program managers and user communities to elicit priorities for T&E benefit from succinct but clear descriptions of potential hazards and precipitating factors; long lists of hazards can quickly become overwhelming and repetitive	**Critical Hazard Scenarios.** Employed the hazard analysis model to identify 6 critical hazard scenarios that can be tailored for specific programs and use contexts, with discussion questions for each scenario to facilitate stakeholder engagement
Testability. Existing risk taxonomies often do not define hazards in a way that enables the development of test strategies, particularly when testers have limited access to models	**Flexible Test Strategies.** Identified multiple test approaches for each hazard scenario to provide options for varying levels of system access and resourcing

Retrieval Augmented Generation-Large Language Model (RAG-LLM) Critical Hazard Scenarios were created and framed in reference model for repeatable use; system under test's specific activity diagrams are then used to form specific T&E approach

Use Case – Intelligence Application of RAG-LLM

This use case includes the use of an AI RAG-LLM to support the intelligence processes described below:

- **Processing and Exploitation**: Convert raw data into usable formats through decryption, translation, filtering, and initial analysis to prepare for deeper evaluation.
- **Analysis and Production**: Evaluating and interpreting processed information to produce actionable intelligence. Analysts assess the reliability, relevance, and significance of the data to create reports, assessments, and forecasts.

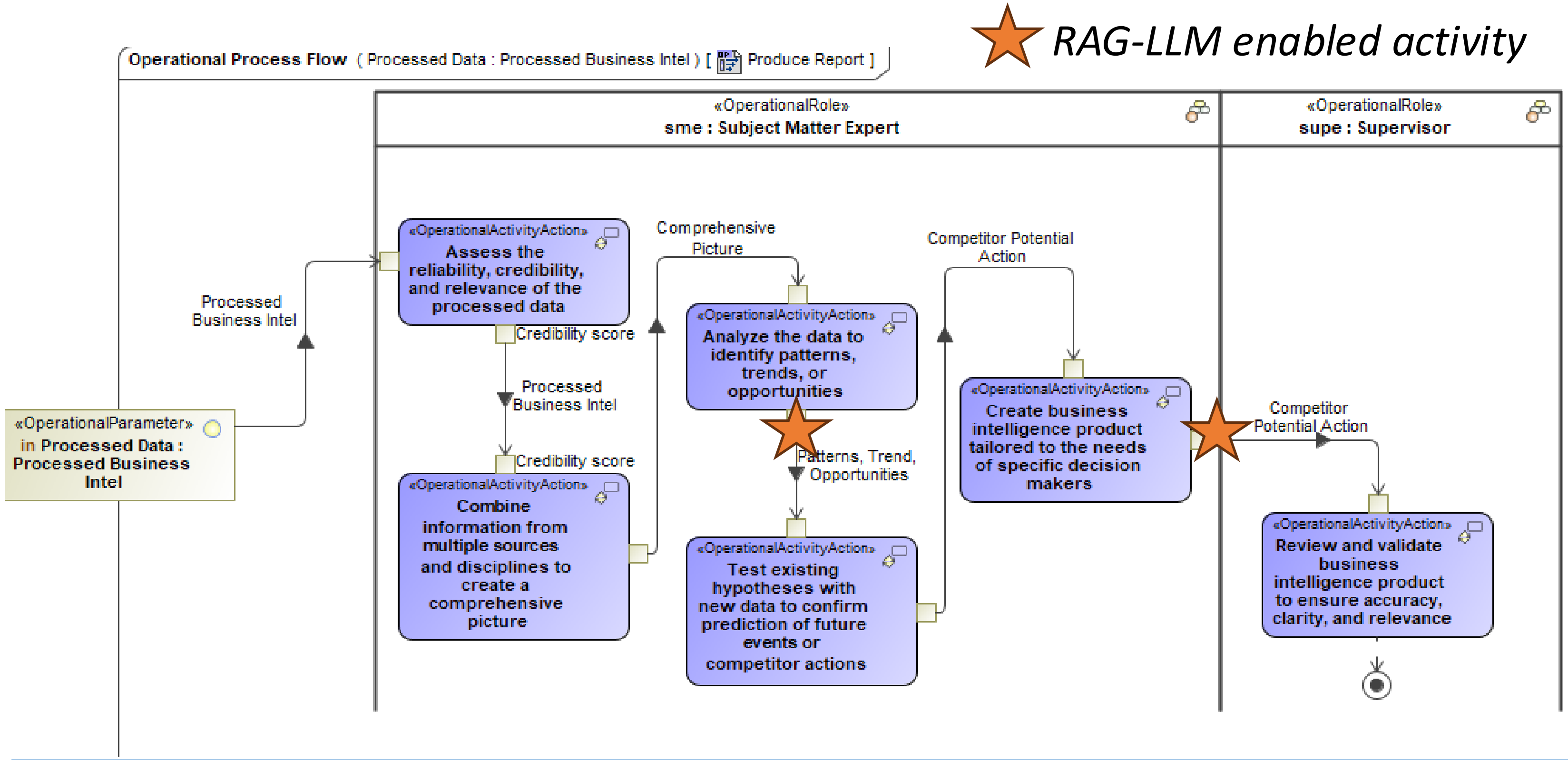


[What Is Business Intelligence (BI)? | IBM]

For this presentation, we will refer to intelligence tasks in the analogous context of business intelligence.

Use Case – Intelligence Application of RAG-LLM

- LLMs and RAGs can serve as interactive tools for analysts and decision-makers, enabling them to ask specific questions and receive detailed, contextually accurate answers about the intelligence they have collected.
- The activity diagram shows which activities will be performed with the RAG-LLM, designated with a star – this allows the team to isolate specific steps to test for hazards tailored to the use case.

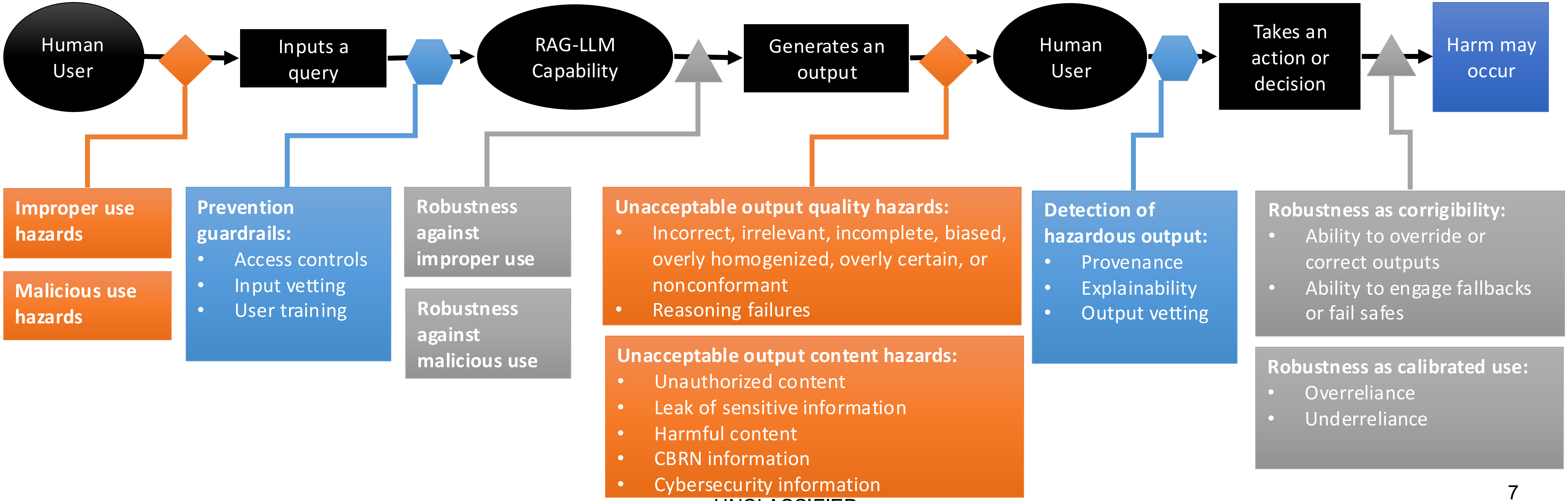


User stories for RAG-LLM enabled activity
As a subject matter expert on business development, I want to use a system to identify industry trends and patterns in competitor activities from the data sources I identify so that I may make recommendations to my supervisor on how my organization will respond to stay competitive.
As a subject matter expert on business development, I want to use a system to quickly answer senior leadership questions with concise and accurate information to enable them to respond rapidly to the environmental changes that affect the commercial space we operate within.

A Hazard Reference Model for RAG-LLM Systems

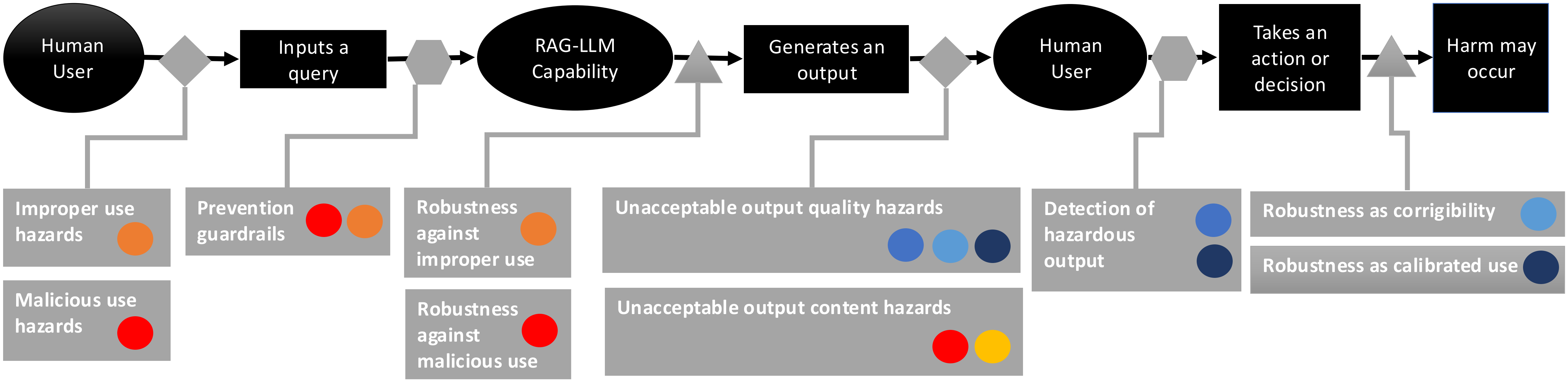
- Human users can inadvertently cause **hazards** because of the open-ended nature of RAG-LLM systems. This same flexibility can also allow bad actors to engage in malicious use. **Guardrails** can be added to the system to mitigate these risks, and AI models can be engineered to be **robust** against misuse.

- RAG-LLM capabilities can also cause **hazards**, either due to poor performance, or because of the nature of how these types of models are trained. **Guardrails** can enable users to detect and correct for these hazards, making the system **robust**.



A Hazard Reference Model for RAG-LLM Systems

The Reference Model specifies how hazards can arise and cascade through the system in six archetypal Critical Hazard Scenarios



- Malicious user succeeds in generating malicious content
- User engages in unauthorized use
- System generates unacceptable output in response to benign prompt
- Unacceptable quality output goes undetected
- User is unable to correct unacceptable quality output
- User is over- or under-reliant on system

UNCLASSIFIED

Prioritization Phase of Hazard Analysis for RAG-LLM BI

- AI SMEs can use the activity model for the system together with the reference model to determine the technical prevalence of hazards based on factors identified using the questions below.
- Users/Mission SMEs then assess the mission impact, and the hazards can be prioritized for testing.

Hazard Scenario	Questions to Consider to Understand Prevalence and Degree of Harm	AI Technical Risk	Mission Impact Severity
User is over- or under-reliant on system	• Does the proposed workflow require explicit human review and approval of system output? • Is there a risk of user skill atrophy at tasks performed using the system? • What might cause a user to fail to adopt the system?	Moderate	Moderate
Malicious user succeeds in generating malicious content	• Is the system intended for use by the public? • Does the system contain sensitive information? • What type of access does the system have to other systems on the network?	High	Moderate
User engages in unauthorized use	• Are there activities in the workflow that must be done by a human, or that are particularly high risk or consequential? • Could the data in the system be used to support other mission activities that must be done by a human, or that are particularly high risk or consequential?	Low	Moderate
Capability generates unacceptable content in response to a benign prompt	• Are the subjects of user inputs likely to touch on sensitive topics? • Might the subjects of user inputs inadvertently trigger system or data biases? • Could it be possible to infer system information from the data in the system?	Low	Low
Unacceptable quality output goes undetected	• Are user inputs likely to be complex reasoning or synthesis tasks? • Is the system data internally contradictory, or contradictory to publicly available information? • Is the system data directly responsive to user queries, or will inference be required? • How will users judge whether output is of acceptable quality?	High	High
User is unable to correct unacceptable quality output	• How does the proposed system allow users to respond to unacceptable outputs? • What fallbacks or fail safes are in place and how to users engage them?	Low	High

Testing Approaches by Hazard Scenario

Hazard Scenario	Level	Testing Approach
Malicious user succeeds in generating malicious content	System	Red teaming to test prevention guardrails, can be done during DT using standard prompts
	Capability	Red teaming to assess robustness to malicious use
	Model	Benchmarks and model cards for malicious content and robustness
User engages in unauthorized use	System	Red teaming to test prevention guardrails
	Capability	Red teaming to assess robustness to unauthorized use
Capability generates unacceptable content in response to a benign prompt	Capability	Red teaming to test specific output hazard
	Model	Benchmarks and model cards for malicious content
Unacceptable quality output goes undetected	System	User testing to test hazard detection guardrails
	Capability	Answer correctness, answer relevance
	Component	Context precision, context recall, context relevance, faithfulness
	Model	Benchmarks and model cards for task performance and hallucinations
User is unable to correct unacceptable quality output	System	User testing to assess corrigibility, fallbacks and fail safes framed within user trust assessments
User is over- or under-reliant on system	System	User testing to detect and account for calibration framed within user trust assessments

AIES Test Levels

Increased Access

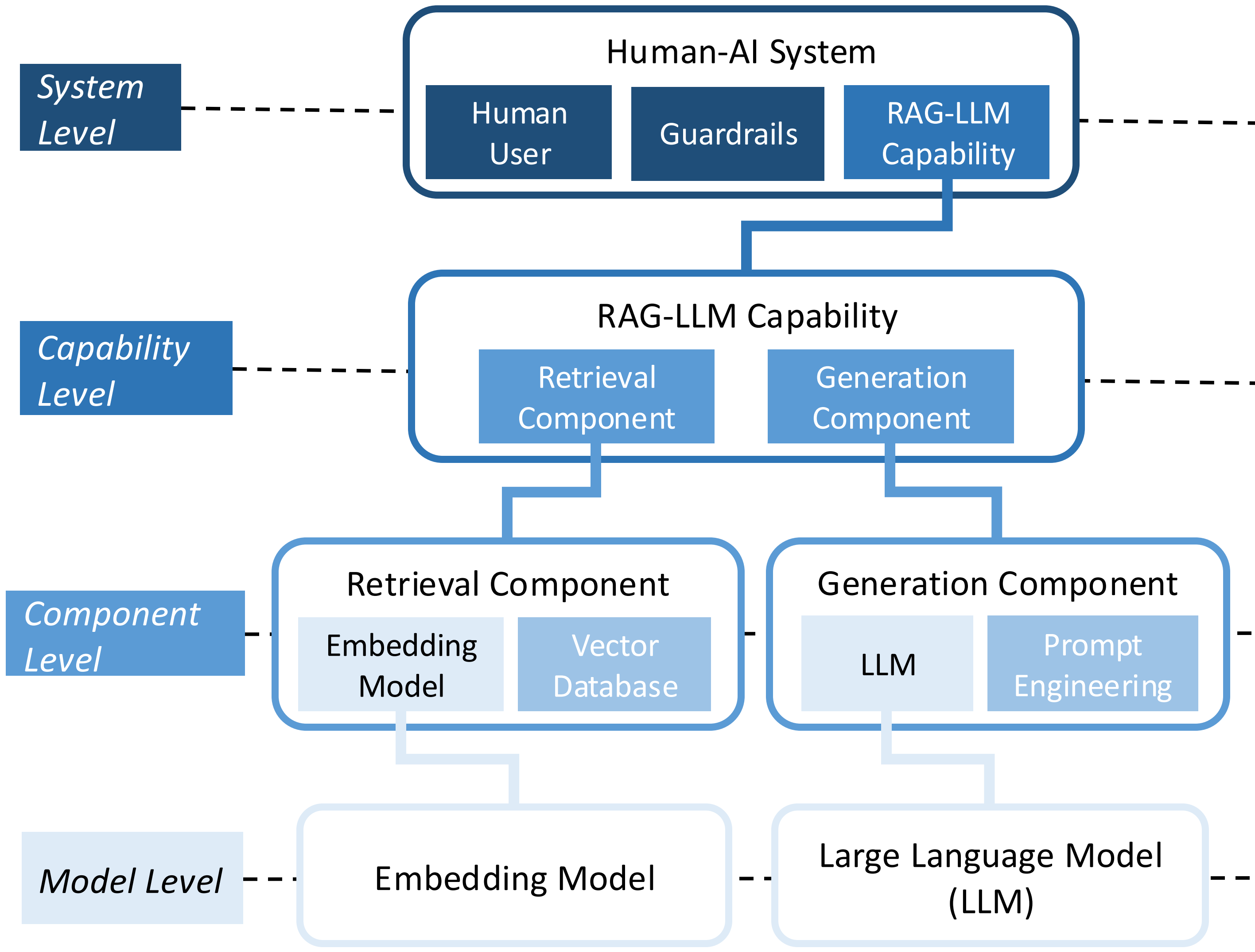
System Level

Capability Level

Component Level

Model Level

Hazard T&E Approaches



T&E Approaches

- Human AI Teaming(HAT) /User testing
- Trust instruments
- Red teaming of guardrails

- Capability performance metrics for output quality
  - Answer correctness and answer relevance
- Red teaming of robustness

- Component performance metrics for output quality
  - Context precision, context recall, context relevance, faithfulness

- Benchmarking
- Model Card review
- Prototyping

BI T&E Approaches

HAT testing to assess hazard handling with/without detection by guardrails, susceptibility to modification, fallbacks and fail safes, framed within user trust assessments

Developmental testing to assess RAG-LLM output during hazard:

- *Answer correctness*
- *Answer relevance*
- *Context precision*
- *Context recall*
- *Context relevance*
- *Faithfulness*
- *Transparency on model limitations*

RAG-LLM – Test edges of LLM context limitations that were intended to be mitigated by RAG

Use standard benchmarks and model cards to assess candidate LLMs for task performance and hallucinations, use available model and data cards to assess candidate embedding models

Summary and Next Steps

- T&E of RAG-LLMs could lead to increases in test complexity.
- We can be more efficient in our use of testing resources (time, tools, SME) by focusing on the likely hazardous scenarios for RAG-LLMs.
- Creating a reference model for hazards for RAG-LLM enables a repeatable process for RAG-LLMs test planning, test execution and reporting.
- The hazards for the System Under Test's are framed within its mission context using MBSE artifacts.
- The team will work to continue piloting these methods and look forward to other communities who may have attempted similar or alternative methods for RAG-LLM evaluations.